Supplementary Material IV: Case Study

This document provides additional details on the meta-analysis case study presented in Section 5 of the article "Meta-Analysis of Generalized Additive Models in Neuroimaging Studies" by Sørensen et al.

We load the following packages:

```
library(metagam)
library(ggplot2)
library(tidyr)
library(dplyr)
library(xtable)
library(purrr)
library(mgcv)
library(latex2exp)
```

The full dataset is contained in a dataframe named `full_data`, while the `cohort_data` is a nested dataframe containing the separate cohorts in a list column, as shown below.

```
cohort_data
```

```
## # A tibble: 6 x 2
## # Groups:   Study [6]
##   Study        data
##   <fct>        <list>
## 1 LCBC         <tibble [2,082 × 34]>
## 2 Cam-CAN      <tibble [866 × 34]>
## 3 BASE-II      <tibble [187 × 34]>
## 4 Betula       <tibble [499 × 34]>
## 5 Whitehall-II <tibble [773 × 34]>
## 6 Barcelona    <tibble [214 × 34]>
```

Separate Fits per Cohort

We first fit one model per cohort.

LCBC

The code below fits the LCBC cohort.

```
LCBC_fit <- gamm(Hippocampus ~ s(Age, k = 20, bs = 'cr') +
                   s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
                 data = cohort_data$data[[1]],
                 random = list(ID =~ 1), method = "REML")
```

Next, we confirm that the number of splines is high enough, using the permutation test described in Wood (2017), Ch. 5.9. If the p-value is non-signficant and the estimated

number of degrees of freedom (edf) is not close to the basis dimension k', we are likely to use a sufficiently large spline basis (number of knots).

```
k.check(LCBC_fit$gam)
```

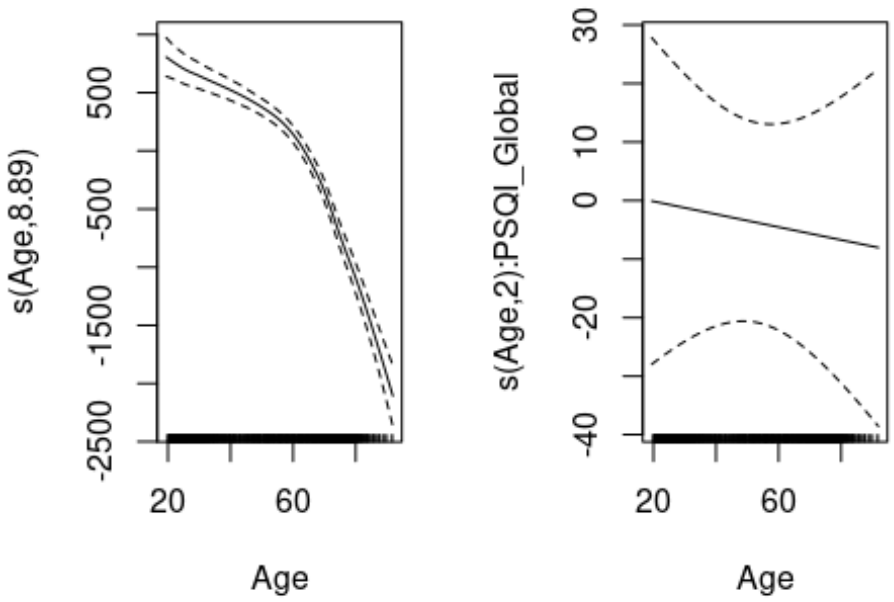
```
##                     k'      edf   k-index p-value
## s(Age)              19 8.892911 0.9897096  0.3150
## s(Age):PSQI_Global   5 2.000015 0.9897096  0.3325
```

We then print out the model summary and plot the smooth terms.

```
summary(LCBC_fit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 20, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7645.54      59.42  128.68   <2e-16 ***
## SexMale       571.41      49.41   11.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F p-value
## s(Age)              8.893  8.893 71.210  <2e-16 ***
## s(Age):PSQI_Global 2.000  2.000  0.155   0.857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.442
##   Scale est. = 15193     n = 2082
```

```
par(mfrow = c(1, 2))
plot(LCBC_fit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



Cam-CAN

The code below fits the Cam-CAN cohort.

```
CamCAN_fit <- gamm(Hippocampus ~ s(Age, k = 10, bs = 'cr') +
                     s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
                   data = cohort_data$data[[2]],
                   random = list(ID =~ 1), method = "REML")
```

Again we confirm that the basis dimension is high enough.

```
k.check(CamCAN_fit$gam)
```

```
##                      k'      edf  k-index p-value
## s(Age)              9 4.134399 1.008887  0.5825
## s(Age):PSQI_Global  5 2.000004 1.008887  0.5850
```

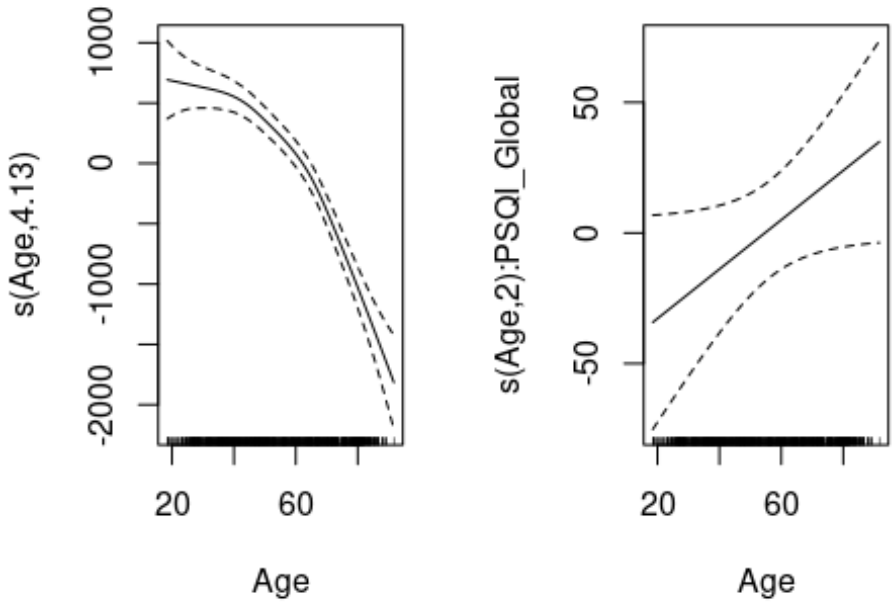
We then print out the model summary and plot the smooth terms.

```
summary(CamCAN_fit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 10, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
```

```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7640.88      64.09  119.22   <2e-16 ***
## SexMale       661.16      58.85   11.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F p-value
## s(Age)              4.134  4.134 42.176  <2e-16 ***
## s(Age):PSQI_Global 2.000  2.000  1.943   0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.416
##   Scale est. = 22722     n = 866
```

```
par(mfrow = c(1, 2))
plot(CamCAN_fit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



BASE-II

Next, a GAMM is computed for the BASE-II data.

```
BASEII_fit <- gamm(Hippocampus ~ s(Age, k = 10, bs = 'cr') +
                     s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
```

```
                  data = cohort_data$data[[3]],
                  random = list(ID =~ 1), method = "REML")
```

Again we confirm that the basis dimension is high enough.

```
k.check(BASEII_fit$gam)
```

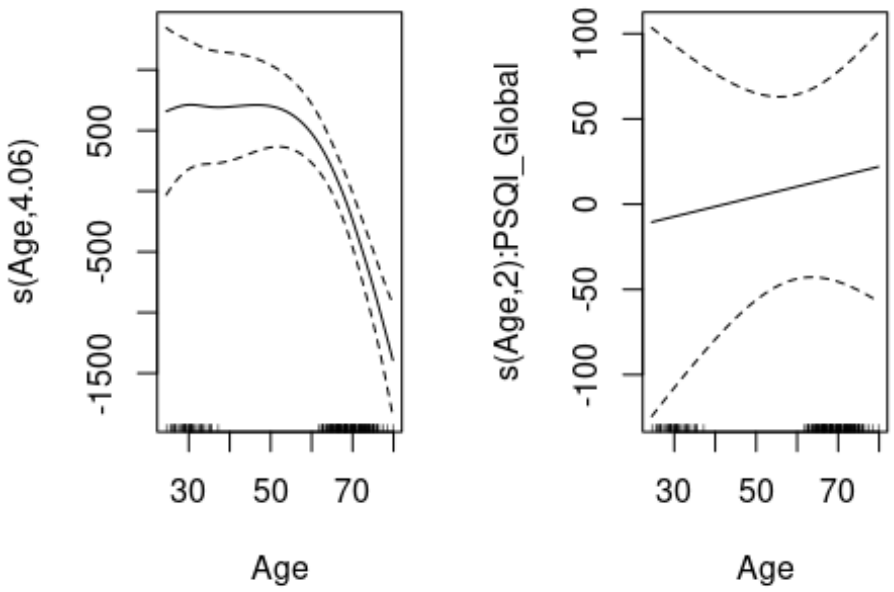
```
##                       k'      edf  k-index p-value
## s(Age)              9 4.062581 1.119666  0.9275
## s(Age):PSQI_Global  5 2.000002 1.119666  0.9350
```

We then print out the model summary and plot the smooth terms.

```
summary(BASEII_fit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 10, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7155.4      177.8  40.253  < 2e-16 ***
## SexMale        704.3      157.3   4.478 1.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F  p-value
## s(Age)              4.063  4.063 16.877 3.84e-12 ***
## s(Age):PSQI_Global 2.000  2.000  0.153    0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.34
##   Scale est. = 13966     n = 187
```

```
par(mfrow = c(1, 2))
plot(BASEII_fit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



Betula

Next, a GAMM is computed for the Betula data.

```
Betula_fit <- gamm(Hippocampus ~ s(Age, k = 10, bs = 'cr') +
                     s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
                   data = cohort_data$data[[4]],
                   random = list(ID =~ 1), method = "REML")
```

Again we confirm that the basis dimension is high enough.

```
k.check(Betula_fit$gam)
```

```
##                      k'      edf   k-index p-value
## s(Age)              9 3.911290 0.9967293  0.4525
## s(Age):PSQI_Global  5 3.671437 0.9967293  0.4825
```

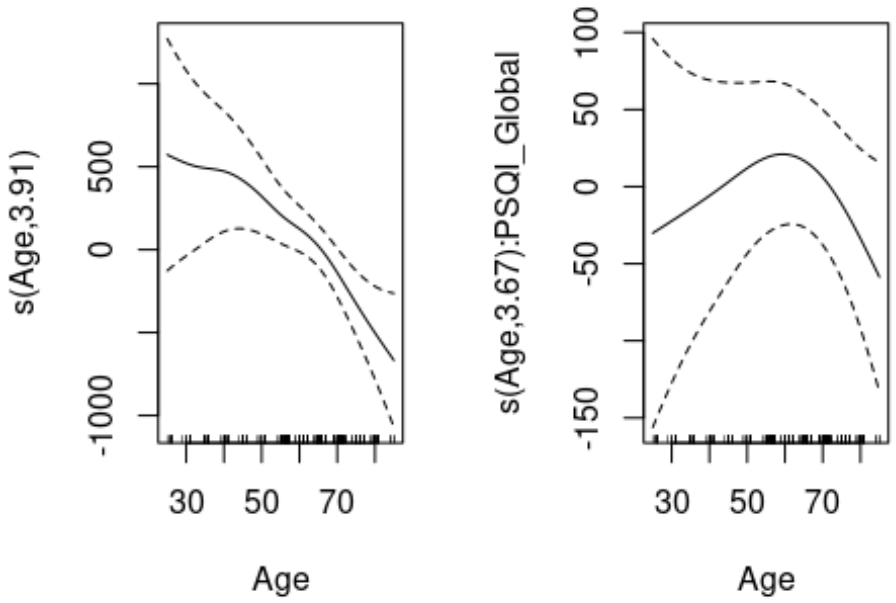
We then print out the model summary and plot the smooth terms.

```
summary(Betula_fit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 10, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
```

```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7612.29     130.30  58.420  < 2e-16 ***
## SexMale       419.53      85.71   4.895 1.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df     F p-value
## s(Age)              3.911  3.911 3.571 0.00761 **
## s(Age):PSQI_Global 3.671  3.671 1.050 0.24593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.323
##   Scale est. = 18930     n = 499
```

```
par(mfrow = c(1, 2))
plot(Betula_fit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



Whitehall-II

Next, a GAM is computed for the Whitehall-II data, since this cohort does not have repeated measurements.

```r
WhitehallII_fit <- gam(Hippocampus ~ s(Age, k = 10, bs = 'cr') +
                        s(Age, by = PSQI_Global, k = 5, bs  = 'cr') + Sex,
                      data = cohort_data$data[[5]], method = "REML")
```

Again we confirm that the basis dimension is high enough.

```r
k.check(WhitehallII_fit)
```

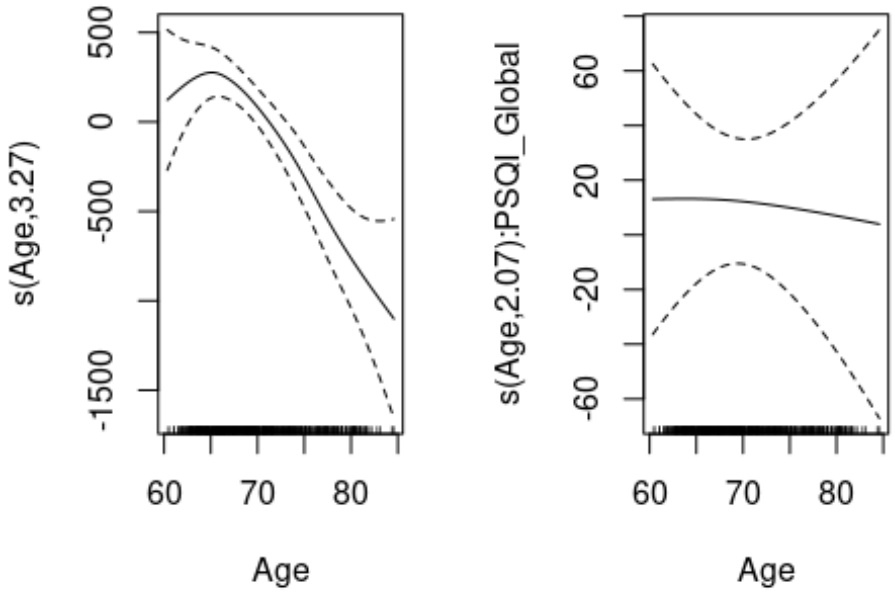
```
##                     k'      edf   k-index p-value
## s(Age)               9 3.267582 0.9767994  0.2750
## s(Age):PSQI_Global   5 2.073277 0.9767994  0.2425
```

We then print out the model summary and plot the smooth terms.

```r
summary(WhitehallII_fit)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 10, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7023.95      93.73  74.940  < 2e-16 ***
## SexMale       336.05      78.14   4.301 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df     F  p-value
## s(Age)              3.268  4.065 8.851 4.45e-07 ***
## s(Age):PSQI_Global 2.073  2.129 0.477    0.601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.153   Deviance explained =   16%
## -REML = 6287.4  Scale est. = 7.2399e+05  n = 773
```

```r
par(mfrow = c(1, 2))
plot(WhitehallII_fit, pages = 1, seWithMean = TRUE, scale = 0)
```



Barcelona

Finally, a GAMM is computed for the Barcelona data.

```
Barcelona_fit <- gamm(Hippocampus ~ s(Age, k = 10, bs = 'cr') +
                        s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
                      data = cohort_data$data[[6]],
                      random = list(ID =~ 1), method = "REML")
```

Again we confirm that the basis dimension is high enough.

```
k.check(Barcelona_fit$gam)
```

```
##                      k'      edf  k-index p-value
## s(Age)              9 2.123726 1.054051  0.7275
## s(Age):PSQI_Global  5 2.000001 1.054051  0.7475
```

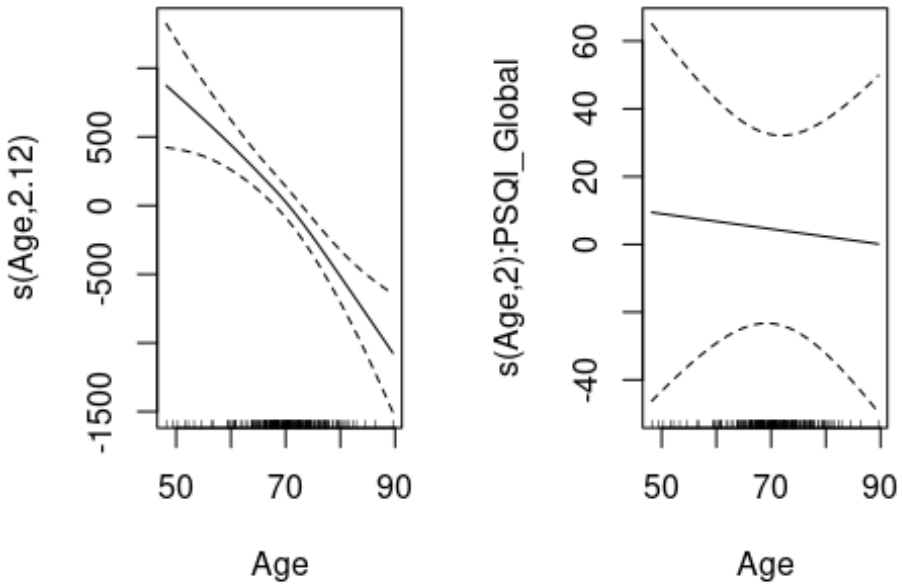
We then print out the model summary and plot the smooth terms.

```
summary(Barcelona_fit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Hippocampus ~ s(Age, k = 10, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
```

```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7238.9      106.1  68.239  < 2e-16 ***
## SexMale        632.1      116.9   5.409 1.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df      F  p-value
## s(Age)              2.124  2.124 20.927 2.43e-09 ***
## s(Age):PSQI_Global 2.000  2.000  0.072    0.931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.29
##   Scale est. = 11847     n = 214
```

```r
par(mfrow = c(1, 2))
plot(Barcelona_fit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



Preparing Meta-Analysis

Next we define a list of models and run the function `strip_rawdata`, which removes all individual participant data from the model objects.

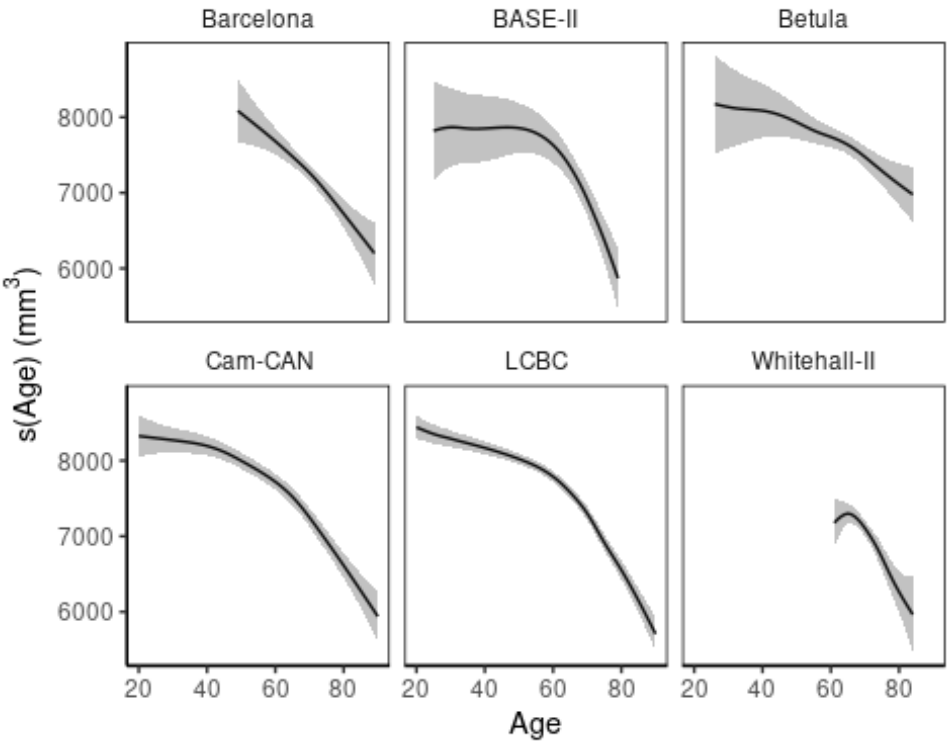
```
fits <- list(Barcelona = Barcelona_fit,
             `BASE-II` = BASEII_fit,
             Betula = Betula_fit,
             `Cam-CAN` = CamCAN_fit,
             LCBC = LCBC_fit,
             `Whitehall-II` = WhitehallII_fit)

cohort_fits <- map(fits, strip_rawdata, save_ranges = TRUE)
```

Then we define the grid over which to compute estimates.

```
grid <- tibble(
  Age = seq(from = 20, to = 90, by = 1),
  Sex = factor("Female", levels = c("Female", "Male")),
  PSQI_Global = 1
)
```

The code below creates a plot of the separate cohort fits, which is also shown in the main paper.

```
pred_data <- imap_dfr(cohort_fits, function(fit, cohort){

  lower <- min(fit$var.summary$Age)
  upper <- max(fit$var.summary$Age)

  grid2 <- grid %>%
    filter(Age > !!lower, Age < !! upper)

  pred <- predict(fit, newdata = grid2, se.fit = TRUE,
                  type = "iterms", terms = c("s(Age)"))

  grid2 %>%
    mutate(
      fit = rowSums(pred$fit) + attr(pred, "constant"),
      se = apply(pred$se.fit, 1, function(x) sqrt(sum(x^2))),
      cohort = cohort
    )
})

p <-  ggplot(pred_data, aes(x = Age, y = fit, ymin = fit + qnorm(.025) * se,
              ymax = fit + qnorm(.975) * se)) +
  geom_line() +
  geom_ribbon(alpha = .3, color = NA) +
  facet_wrap(vars(cohort)) +
  theme_classic() +
  theme(strip.background = element_blank(),
        panel.border = element_rect(colour = "black", fill = NA)) +
  ylab(TeX("s(Age) (mm$^{3}$)"))

p
```



Mega-Analysis

The code below fits a GAMM to the complete dataset. Note that we use nested random effects, specified by `list(Study =~ 1, ID =~ 1)`. The upper level is cohort (variable `Study`) within which individual subjects are nested (variable `ID`).

```
fullfit <- gamm(Hippocampus ~ s(Age, k = 20, bs = 'cr') +
                  s(Age, by = PSQI_Global, k = 5, bs = 'cr') + Sex,
                data = full_data,
                random = list(Study =~ 1, ID =~ 1), method = "REML")
```

Again it seems that we are using sufficiently many splines.

```
k.check(fullfit$gam)
```

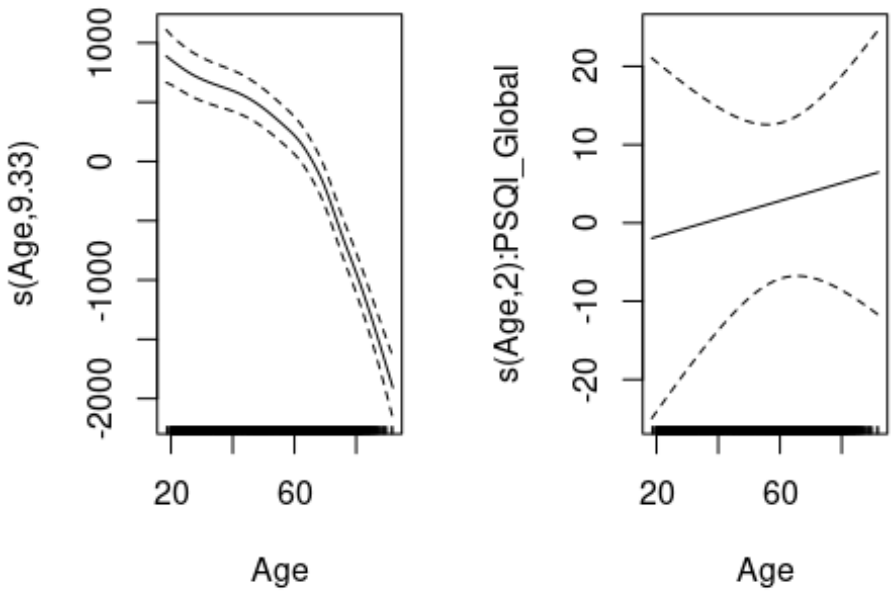
```
##                     k'      edf  k-index p-value
## s(Age)              19 9.333934 1.005233  0.6475
## s(Age):PSQI_Global   5 2.000030 1.005233  0.6475
```

We print some model output and plot the estimated smooth terms.

```
summary(fullfit$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## Hippocampus ~ s(Age, k = 20, bs = "cr") + s(Age, by = PSQI_Global,
##     k = 5, bs = "cr") + Sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7470.24      83.65   89.31   <2e-16 ***
## SexMale       522.76      30.73   17.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df       F p-value
## s(Age)              9.334  9.334 106.131  <2e-16 ***
## s(Age):PSQI_Global 2.000  2.000   0.279   0.757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.372
##   Scale est. = 16960     n = 4621

plot(fullfit$gam, pages = 1, seWithMean = TRUE, scale = 0)
```



Meta-Analysis

We conduct analysis over the following grid. Note that `mgcv` requires all variables to be present in the grid, even when predicting single terms.

```
grid <- tibble(
  Age = seq(from = 20, to = 90, by = .1),
  Sex = factor("Female", levels = c("Female", "Male")), PSQI_Global = 1
)
```

Main Effect of Age

The following computes the term `s(Age)` for the mega-analysis over the grid defined above.

```
pred <- predict(fullfit$gam, newdata = grid,
                se.fit = TRUE, type = "iterms",
                terms = "s(Age)", newdata.guaranteed = TRUE)

fullpred <- grid %>%
  mutate(
    estimate = c(pred$fit) + attr(pred, "constant"),
    se = c(pred$se.fit)
  )
```

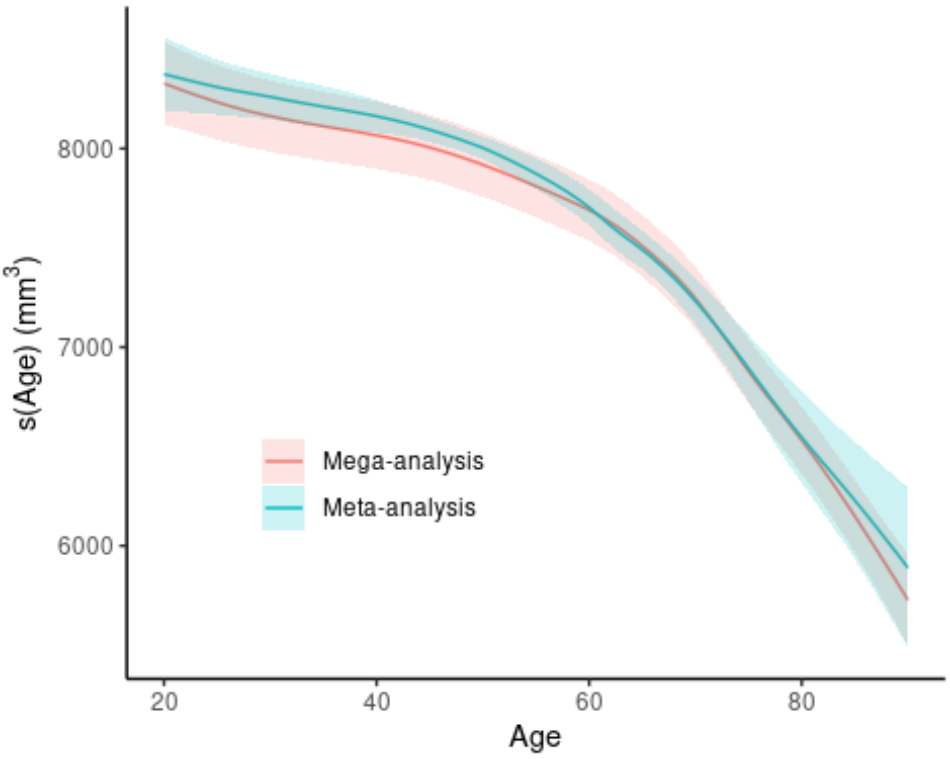
The following computes a meta-analytic fit over the same grid, using random-effects meta-analysis with the method-of-moments estimator.

```
metafit <- metagam(cohort_fits, grid, type = "iterms",
                   terms = "s(Age)", method = "DL", intercept = TRUE)
```

The following computes a plot comparing the meta- and mega-analytic fits. This plot is also shown in the main paper.

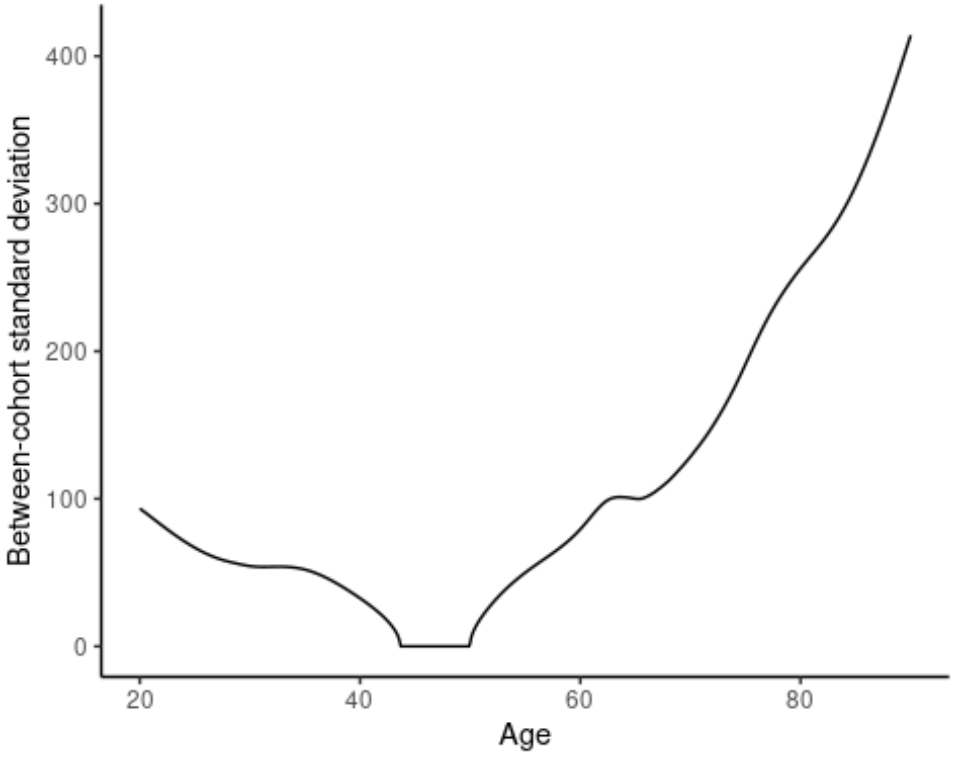
```
plot_df <- bind_rows(
  meta = metafit$meta_estimates,
  full = fullpred,
  .id = "type"
) %>%
  mutate(type = if_else(type == "full", "Mega-analysis", "Meta-analysis"))

p1 <- ggplot(plot_df, aes(x = Age, y = estimate, ymin = estimate + qnorm(.025
) * se,
                                ymax = estimate + qnorm(.975) * se,
              group = type, color = type, fill = type)) +
  geom_line() +
  geom_ribbon(alpha = .2, color = NA) +
  theme_classic() +
  theme(legend.title = element_blank(), legend.position = c(.3, .3)) +
  ylab(TeX("s(Age) (mm$^{3}$)"))
```

```
p1
```



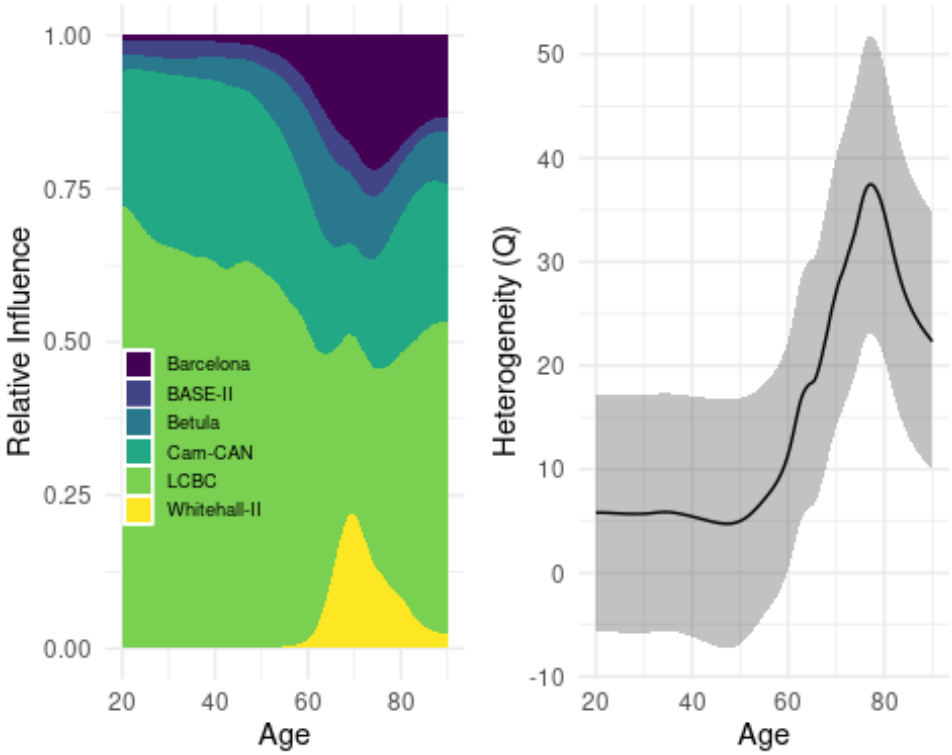
The code below extracts and plots the estimated between-study standard deviation from the estimated model object.

```
tibble(
  Between_sd = map_dbl(metafit$meta_estimates$meta_model, ~ sqrt(.x$tau2)),
  Age = metafit$meta_estimates$Age
) %>%
  ggplot(aes(x = Age, y = Between_sd)) +
  geom_line() +
  theme_classic() +
  ylab("Between-cohort standard deviation")
```



Finally we make a dominance plot and a heterogeneity plot, both of which are shown in the main paper.

```
pd1 <- plot_dominance(metafit) +
  theme(legend.position = c(.25, .37),
        legend.key = element_rect(color = "white"),
        legend.title = element_blank(),
        legend.text = element_text(size = 7),
        legend.key.size = unit(.75, "line"))
pd2 <- plot_heterogeneity(metafit)
p3 <- cowplot::plot_grid(pd1, pd2)

p3
```



Interaction Term

Next, we consider the varying-coefficient term.

First, we compute the mega-analytic fit.

```
pred <- predict(fullfit$gam, newdata = grid, se.fit = TRUE,
                type = "iterms", terms = "s(Age):PSQI_Global")

fullpred <- grid %>%
  mutate(
    estimate = rowSums(pred$fit),
    se = pred$se.fit
  )
```

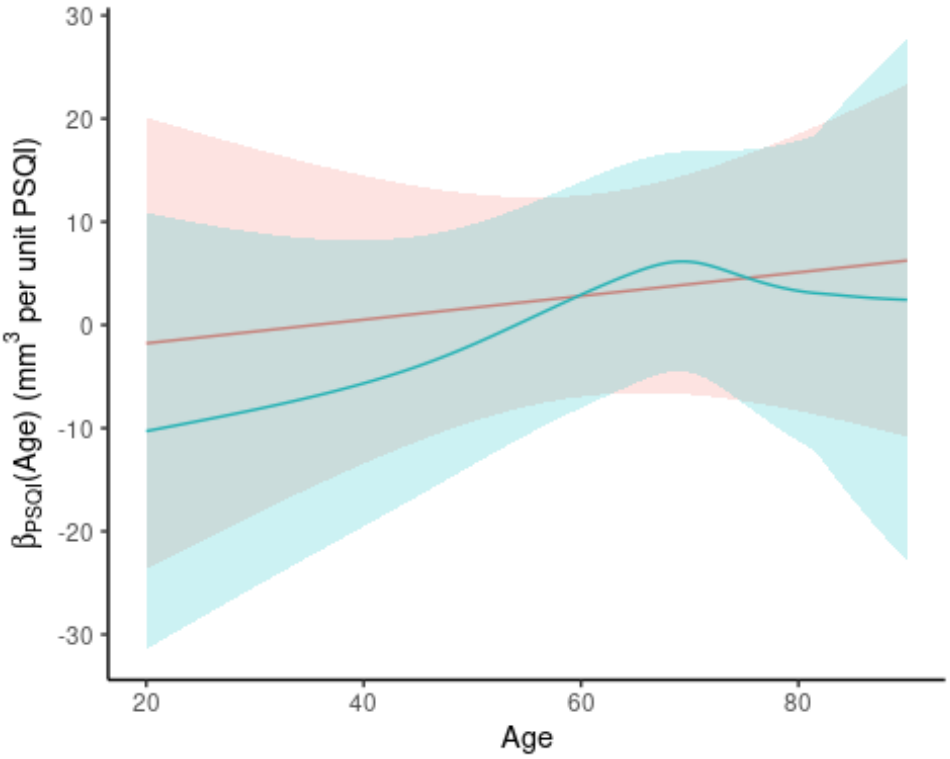
The we compute the meta-analytic fit, using `metagam`.

```
metafit <- metagam(cohort_fits, grid, type = "iterms", terms = "s(Age):PSQI_G
lobal",
                   method  = "DL", intercept = FALSE)
```

The code below creates a plot comparing the terms, as also shown in the paper.

```
p2 <- bind_rows(
  meta = metafit$meta_estimates,
  full = fullpred,
  .id = "type"
) %>%
```

```r
  mutate(type = if_else(type == "full", "Mega-analysis", "Meta-analysis")) %>
%
  ggplot(aes(x = Age, y = estimate, ymin = estimate + qnorm(.025) * se,
                                    ymax = estimate + qnorm(.975) * se,
             group = type, color = type, fill = type)) +
  geom_line() +
  geom_ribbon(alpha = .2, color = NA) +
  theme_classic() +
  theme(legend.position = "none") +
  ylab(TeX("$\\beta_{PSQI}(Age)$ (mm$^{3}$ per unit PSQI)"))


p2
```



References

Wood, S.N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman; Hall/CRC.