Analysis protocol of the "Single-centre study reports 84% five-year overall survival rate for all solid paediatric tumours" -study

Eero Teppo

Last update 2016-01-24

1. Software

This document and analyses were done using R version 3.2.2 with "x86_64-w64-mingw32" platform, with RStudio, RMarkdown, and several external packages.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

2. Research questions guiding analyses

Batch 1. Survival experience inference:

- What was the 1-year and 5-year overall survival probability of...
  - a. ICCC3 group III (CNS) patients?
  - b. non-III patients
  - c. all solid tumor patients?
- What was the survival function of...
  - a. all patients by ICCC3 main tumor classes?
  - b. the ICCC3 group II patients (lymphomas) by II-subclasses?
  - c. non-III patients by transplantation?
  - d. III patients by transplantation?
- How this data compares to European and Finnish estimates? (Gatta et al. and Madanat-Harjuoja et al.)
- How different was survival in patients diagnosed in 1990-1999 and 2000-2014?

Survival experience comparison inference:

- Was there a difference between survival functions of...
  - a. patient populations with different stage at diagnosis?
  - b. neuroblastoma patient populations with different stage at diagnosis?

Batch 2. Stage distribution inference:

- What was the stage distribution of...
  - a. all patients by ICCC3 main tumor classes?

Batch 3. Tumor size distribution comparison inference:

- Was there a difference between tumor size distributions of
  - a. patient populations with different stage at diagnosis?
  - b. III and non-III groups of patients?
- What was the difference between tumor size distributions of males and females in III and non-III patient populations?
- What was the correlation of tumor size and age at diagnosis in III and non-III patient populations?

Batch 4. Transplantation proportion inference:

- What proportion of patients by ICCC3 -main groups was treated with autologous stem cell transplantation?

Batch 5. Diagnostic imaging trend inference:

- What were the trends of diagnostic imaging modalities like?
- How likely was there a change in the populations of trends of diagnostic imaging modalities?

Batch 6. Metastasis proportion inference:

- What was the proportion of metastasised cancers in...
  - a. III group?
  - b. non-III group?

Batch 7. Exploratory Cox regression modeling of survival time and predictors

Batch 8. The correlation of body area and tumor size:

- What was the correlation of tumor size and body area in non-CNS patients?

3. Raw data cleaning

3.1. Reading

```
path_raw <- "G:/Projects/potti/data/raw/potti-kk-20150529.xlsx"
library(readxl)
pottikk <- read_excel(path = path_raw)
```

3.2. Coercing to tdb_df-class

```
library(dplyr)
pottikk <- tbl_df(pottikk)
```

3.3. Combining, selecting, filtering, and renaming

- Only relevant variables (columns) and observations (rows) are selected from the raw data.

```
# necessary class and name changes
require(lubridate)
pottikk$Syntymäaika <- dmy(pottikk$Syntymäaika)
names(pottikk) <- gsub(" ", "_", names(pottikk))
names(pottikk) <- gsub("-", "_", names(pottikk))
pottikk <- pottikk %>%
  # selecting variables concerning research questions
  select(id = Id,
         birth_date = Syntymäaika,
         death_date = Kuolinpvm,
         sex1 = Sukupuoli,
         ds = Ds,
         iccc_main = ICCC3_paaluokka_ICCC3_luokitus,
         iccc_sub = ICCC3_alaluokka_ICCC3_luokitus,
         dx_date = DiagnoosiPvm_KKdiagnoosivaihe,
         tumorsize = Primaarikasvaimenkok_KKdiagnoosivaihe,
         stage = Stage_KKdiagnoosivaihe,
         sex2 = Sukupuoli_KKdiagnoosivaihe,
         hsct_date = Kantasolusiirtopvm_KKkantasolusiirto,
         hsct_type = Siirretyyppi_KKkantasolusiirto,
         hsct_place = Siirtopaikka_KKkantasolusiirto,
         bonemap = Luustokartta_KKkuvantaminenDiag,
         mri_local = MagneettiPaikallinen_KKkuvantaminenDiag,
         mibg = MIBG_KKkuvantaminenDiag,
         mri = MRI_KKkuvantaminenDiag,
         pet = PET_CT_TT_KKkuvantaminenDiag,
         ct_lung = TTkeuhkot_KKkuvantaminenDiag,
         ct_local = TTpaikallinen_KKkuvantaminenDiag,
         primary = Syopatauti_Syopatauti) %>%
  # filtering a) primary cases b) aged 0 - 18 years at diagnosis
  filter(ds == 1) %>%
  mutate(age_dx = as.period(dx_date - birth_date) / years(1)) %>%
  filter(age_dx <= 18 & age_dx >= 0)
```

3.4. Fixing existing, making new and dropping old variables

a) Making new variables

```
# censoring date at the last database update
pottikk <- pottikk %>%
  mutate(censor_date = ymd("2015-05-29"))
```

b) Infering new variables

```
pottikk <- pottikk %>%
  # death indicator from date data
  mutate(death = ifelse(is.na(death_date), 0, 1)) %>%
  # hsct indicator from date or type or place data (note!)
```

```
  mutate(hsct = ifelse(!is.na(hsct_date) |
                         !is.na(hsct_type) |
                         !is.na(hsct_place),
                yes = 1,
                no = 0)) %>%
  # follow-up time from diagnosis date
  mutate(followup =
           ifelse(death == 1,
                  as.period(death_date - dx_date) / years(1),
                  as.period(censor_date - dx_date) / years(1))) %>%
  # year of birth, diagnosis and death variables
  mutate(birth_year = year(birth_date),
         death_year = year(death_date),
         dx_year = year(dx_date)) %>%
  # iccc groups
  mutate(iccc = paste(iccc_main, iccc_sub, sep="")) %>%
  mutate(iccc = ifelse(iccc == "NANA", NA, iccc)) %>%
  # new imaging modality variables
  mutate(bonemap = ifelse(bonemap == "Ei tehty", 0, 1),
         mri = ifelse(mri_local == "Ei" & mri == "Ei", 0, 1),
         mibg = ifelse(mibg == "Ei tehty", 0, 1),
         pet = ifelse(pet == "Ei", 0, 1),
         ct = ifelse(ct_lung == "Ei" & ct_local == "Ei", 0, 1)) %>%
  # drop unnecessary variables
  select(-censor_date, -hsct_type, -hsct_date,
         -dx_date, -ds, -birth_date, -death_date,
         -iccc_sub, -ct_lung, -ct_local, -mri_local)
```

c) Categorizing

```
require(stringr)
pottikk <- pottikk %>%
  # stage to 1, 2, 3, 4, 4S, 5 and M
  mutate(stage = ifelse(stage == "4s" |
                          stage == "5" |
                          stage == "M",
                        stage,
                        str_extract(stage, "\\d"))) %>%
  # stage 5 and M to stage 4
  mutate(stage = ifelse(stage == "5" | stage == "M",
                        yes = "4",
                        no = stage)) %>%
  # iccc III and others -variable
  mutate(iccc_cns = ifelse(iccc_main == "III",
                           yes = 1,
                           no = ifelse(is.na(iccc_main), NA, 0))) %>%
  # categorising too continuous variables for frequency tabulation
  mutate(death_year_cat = cut(death_year,
                              breaks=c(1994, 1999, 2004, 2009, 2015),
                              labels=c("1995-2000", "2000-2005",
                                       "2005-2010", "2010-2015*")),
         followup_cat = cut(followup,
                            breaks=c(-1, 4, 8, 12, 16, 20, 24, 28),
```

```
                            labels=c("0-4", "4-8", "8-12", "12-16",
                                     "16-20", "20-24", "24-28")),
         birth_year_cat = cut(birth_year,
                              breaks=c(1979, 1990, 2001, 2013),
                              labels=c("1980-1990", "1991-2001", "2002-2013")),
         dx_year_cat = cut(dx_year,
                           breaks=c(1986, 1989, 1994, 1999, 2004, 2009, 2015),
                           labels=c("1987-1990", "1990-1995", "1995-2000",
                                    "2000-2005", "2005-2010", "2010-2015*")),
         age_dx_cat = cut(age_dx,
                          breaks=c(-1, 6, 12, 18),
                          labels=c("0-6","6-12", "12-18")),
         tumorsize_cat = cut(tumorsize,
                             breaks=c(0, 4.9999, 9.9999, 14.9999, 20),
                             labels=c("0-5", "5-10", "10-15", "15-20")))
  # checking reveals NA's introduced in categorization
apply(pottikk, 2, function(x) {sum(is.na(x))})
```

```
##             id           sex1      iccc_main      tumorsize          stage
##              0              6             30            136             80
##           sex2     hsct_place        bonemap           mibg            mri
##              9            394             32             27             19
##            pet        primary         age_dx          death           hsct
##             26              3              0              0              0
##       followup     birth_year     death_year        dx_year           iccc
##              0              0            376              0             30
##             ct       iccc_cns death_year_cat   followup_cat birth_year_cat
##             25             30            376              0              0
##    dx_year_cat     age_dx_cat  tumorsize_cat
##              0              0            136
```

d) Fixing

```
pottikk <- pottikk %>%
  # sex variable integration
  mutate(sex = ifelse(is.na(sex1),
                      sex2,
                      sex1)) %>%
  # drop sex1 and sex2
  select(-sex1, -sex2) %>%
  # age less than 0 at diagnosis to NA
  mutate(age_dx = ifelse(age_dx < 0,
                         NA,
                         age_dx)) %>%
  mutate(sex = ifelse(sex == "Mies", "Boy", "Girl"))

# iccc-information missing -> "missing" (for tabulation)
pottikk$iccc_main[is.na(pottikk$iccc_main)] <- "Missing"
```

3.5. Checking classes

Characters, numerics, and dates

```r
# factors to characters
factors <- sapply(pottikk, is.factor)
pottikk[factors] <- lapply(pottikk[factors], as.character)
# id to character
pottikk$id <- as.character(pottikk$id)
rm(factors)
```

3.6. Arranging observations

```r
pottikk <- pottikk %>%
  select(id, primary, iccc_cns, iccc_main, iccc,
         death, death_year, death_year_cat, followup,
         followup_cat, birth_year, birth_year_cat, dx_year,
         dx_year_cat, sex, age_dx, age_dx_cat, stage, tumorsize,
         tumorsize_cat, hsct, hsct_place, bonemap,
         mibg, pet, ct, mri) %>%
  arrange(id)
```

3.7. Describing the dataset

Frequency tabulation (Table 1 for manuscript)

```r
## Frequency tabulation of categorized variables
library(tidyr)
table1 <- pottikk %>%
  # Select categorical variables in the order of tabulation
  select(id, iccc_main, birth_year_cat, sex, age_dx_cat, primary, stage, tumorsize_cat,
         hsct, hsct_place, death, death_year_cat, bonemap, mibg, pet, ct, mri) %>%
  # To longest format
  gather("variable", "value", -id, -iccc_main) %>%
  # Summarise values to category frequencies
  group_by(iccc_main, variable, value) %>%
  summarise(n = n()) %>%
  # Setup for iccc3 column spread
  ungroup() %>%
  mutate(row = 1:nrow(.)) %>%
  # ICCC3 to columns
  spread(iccc_main, n) %>%
  # Summarise again to frequencies
  group_by(variable, value) %>%
  summarise_each(funs(sum(.,na.rm=TRUE))) %>%
  # Clean up
  ungroup() %>%
  select(-row) %>%
  # Row total -column
  mutate(all = Missing + II + III + IV + V +
           VI + VII + VIII + IX + X + XI + XII) %>%
  # Order columns
  select(variable, value, all, II, III, IV,
         V, VI, VII, VIII, IX, X, XI, XII, Missing)
```

```
# To table with printr -package
library(printr)
table1
```

variable	value	all	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Missing
birth_year_cat	1980-1990	84	24	21	2	2	9	0	4	8	2	6	0	6
birth_year_cat	1991-2001	220	39	81	14	4	12	4	8	21	9	9	2	17
birth_year_cat	2002-2013	143	19	50	21	6	17	2	1	13	5	0	2	7
sex	Boy	263	57	91	20	12	24	4	7	21	7	4	1	15
sex	Girl	184	25	61	17	0	14	2	6	21	9	11	3	15
age_dx_cat	0-6	207	17	75	32	11	30	4	0	16	8	0	4	10
age_dx_cat	12-18	119	35	30	1	1	3	1	9	13	4	12	0	10
age_dx_cat	6-12	121	30	47	4	0	5	1	4	13	4	3	0	10
primary	1	435	81	151	37	12	38	6	12	41	16	15	4	22
primary	2	9	0	0	0	0	0	0	1	0	0	0	0	8
primary	NA	3	1	1	0	0	0	0	0	1	0	0	0	0
stage	1	172	18	73	9	7	7	2	4	23	6	3	2	18
stage	2	68	22	20	7	0	13	0	0	1	3	1	0	1
stage	3	50	16	16	1	0	4	0	1	3	2	4	1	2
stage	4	77	14	12	18	1	10	3	5	6	3	4	0	1
stage	NA	80	12	31	2	4	4	1	3	9	2	3	1	8
tumorsize_cat	0-5	146	23	54	18	1	4	0	2	16	7	10	1	10
tumorsize_cat	10-15	58	13	1	6	0	17	2	2	7	6	1	1	2
tumorsize_cat	15-20	22	8	0	2	0	3	2	0	4	1	0	0	2
tumorsize_cat	5-10	85	14	31	8	0	10	2	5	6	2	1	1	5
tumorsize_cat	NA	136	24	66	3	11	4	0	4	9	0	3	1	11
hsct	0	394	76	145	25	10	32	6	10	30	14	15	3	28
hsct	1	53	6	7	12	2	6	0	3	12	2	0	1	2
hsct_place	HUS	6	2	1	1	1	1	0	0	0	0	0	0	0
hsct_place	TAYS	47	4	6	11	1	5	0	3	12	2	0	1	2
hsct_place	NA	394	76	145	25	10	32	6	10	30	14	15	3	28
death	0	376	74	121	30	11	33	2	9	36	15	13	4	28
death	1	71	8	31	7	1	5	4	4	6	1	2	0	2
death_year_cat	1995-2000	8	2	2	0	1	1	0	1	0	0	0	0	1
death_year_cat	2000-2005	18	3	8	2	0	1	1	0	2	0	0	0	1
death_year_cat	2005-2010	23	2	8	3	0	3	2	0	2	1	2	0	0
death_year_cat	2010-2015*	22	1	13	2	0	0	1	3	2	0	0	0	0
death_year_cat	NA	376	74	121	30	11	33	2	9	36	15	13	4	28
bonemap	0	260	45	130	12	9	19	3	0	11	5	8	3	15
bonemap	1	155	33	11	23	0	16	3	12	31	11	5	1	9
bonemap	NA	32	4	11	2	3	3	0	1	0	0	2	0	6
mibg	0	361	75	133	2	10	30	5	12	39	15	12	4	24
mibg	1	59	4	8	33	1	5	1	1	2	1	1	0	2
mibg	NA	27	3	11	2	1	3	0	0	1	0	2	0	4
pet	0	393	63	141	35	11	34	6	10	41	15	7	4	26
pet	1	28	15	0	0	0	2	0	3	1	1	6	0	0
pet	NA	26	4	11	2	1	2	0	0	0	0	2	0	4
ct	0	158	22	76	11	7	6	0	2	11	7	5	0	11
ct	1	264	57	65	24	4	30	6	11	31	9	8	4	15
ct	NA	25	3	11	2	1	2	0	0	0	0	2	0	4
mri	0	78	27	6	7	4	18	1	1	6	3	2	0	3
mri	1	350	53	138	28	7	18	5	12	36	13	11	4	25
mri	NA	19	2	8	2	1	2	0	0	0	0	2	0	2

variable	value	all	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Missing

```
# Delete and fix tabulation-only variables
pottikk <- pottikk %>%
  select(-death_year_cat, -followup_cat, -birth_year_cat,
         -dx_year_cat, -age_dx_cat, -tumorsize_cat)
pottikk$iccc_main <- gsub("Missing", NA, pottikk$iccc_main)
```

3.8. Saving clean study dataset

```
path_clean <- "G:/TUTKIMUS/potti/data/final-tidy"
# setwd(path_clean)
save(pottikk, file="potti-kk-20150529-clean.RData")
# for (i in 1:2) setwd("../")
```

4. Batch 1: Patient populations' survival experience analyses

Nine questions regarding survival experience and their comparisons will be analysed in this first batch.

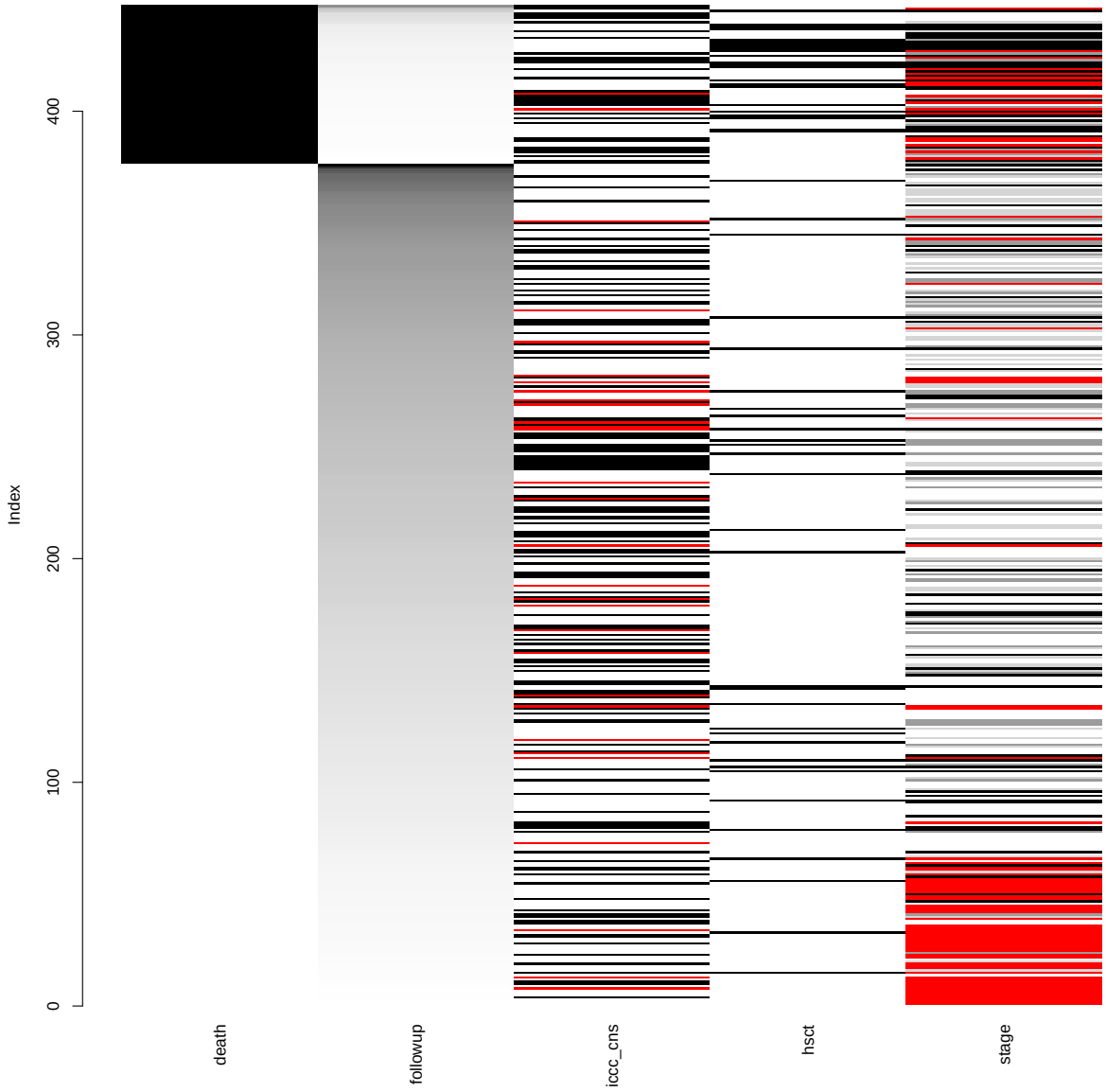
4.1. Selecting and filtering from clean data

```
library(dplyr)
## Variables needed:
## iccc variables, death indicator, follow-up, stage, hsct, hsct place, diagnosis year
surv_exp <- pottikk %>%
  select(death, followup, iccc_cns, iccc_main,
         iccc, hsct, hsct_place, stage, dx_year)
```

4.2. Describing and Handling missing values

Missingness structure

```
library(VIM)
library(dplyr)
matrixplot(arrange(surv_exp[ , c(1, 2, 3, 6, 8)], death, followup))
```



Handling: Complete case analyses

4.3. Main analysis

Primary analytics methods: Kaplan-Meier (95% CI, log), log-rank test (no weight)

```
# Fit kaplan-meier curves as in questions; 9 fits in total
library(survival)
library(dplyr)
# Empty list for fit results
surv_exp_fits <- list()
## 1. All
surv_exp_fits[[1]] <- with(surv_exp,
```

```
                    survfit(formula = Surv(time=followup, event=death) ~ 1))
## 2. III
surv_exp_fits[[2]] <- with(filter(surv_exp, iccc_main=="III"),
                    survfit(formula = Surv(time=followup, event=death) ~ 1))
## 3. non-III
surv_exp_fits[[3]] <- with(filter(surv_exp, iccc_main != "III" & !is.na(iccc_main)),
                    survfit(formula = Surv(time=followup, event=death) ~ 1))
## 4. ICCC3 main classes
surv_exp_fits[[4]] <- with(filter(surv_exp, !is.na(iccc_main)),
                    survfit(formula = Surv(time=followup, event=death) ~ iccc_main))
## 5. III by transplantation
surv_exp_fits[[5]] <- with(filter(surv_exp, iccc_main == "III" &
                                  (hsct_place == "TAYS" |
                                  is.na(hsct_place))),
                    survfit(formula = Surv(time=followup, event=death) ~ hsct))
## 6. non-III by transplantation
surv_exp_fits[[6]] <- with(filter(surv_exp, iccc_main != "III" &
                                  !is.na(iccc_main) &
                                  (hsct_place == "TAYS" |
                                  is.na(hsct_place))),
                    survfit(formula = Surv(time=followup, event=death) ~ hsct))
## 7. II by sub-classes
surv_exp_fits[[7]] <- with(filter(surv_exp, iccc_main == "II"),
                    survfit(formula = Surv(time=followup, event=death) ~ iccc))
# 8. IV by stage
## a. fit for visualization
surv_exp_fits[[8]] <- with(filter(surv_exp, iccc_main == "IV" & !is.na(stage)),
                    survfit(formula = Surv(time=followup, event=death) ~ stage))
# 9. All by stage
## a. fit for visualization
surv_exp_fits[[9]] <- with(filter(surv_exp, !is.na(stage)),
                    survfit(formula = Surv(time=followup, event=death) ~ stage))
```

```
library(survival)
library(dplyr)
# Empty list for comparison results
surv_exp_comp <- list()
# 8. IV by stage
surv_exp_comp[[1]] <- with(filter(surv_exp, iccc_main == "IV" & !is.na(stage)),
                           survdiff(formula = Surv(time=followup, event=death) ~ stage))
# 9. All by stage
surv_exp_comp[[2]] <- with(filter(surv_exp, !is.na(stage)),
                           survdiff(formula = Surv(time=followup, event=death) ~ stage))
# 5. Non-CNS by transplantation
surv_exp_comp[[3]] <- with(filter(surv_exp, iccc_main != "III" &
                                  !is.na(iccc_main) &
                                  (hsct_place == "TAYS" |
                                  is.na(hsct_place))),
                           survdiff(formula = Surv(time=followup, event=death) ~ hsct))

# 6. CNS (III) by transplantation
surv_exp_comp[[4]] <- with(filter(surv_exp, iccc_main == "III" &
                                  (hsct_place == "TAYS" |
```

```
                                    is.na(hsct_place))),
                           survdiff(formula = Surv(time=followup, event=death) ~ hsct))
```

4.3.1. Investigating the assumption of similar prospects: How different was survival in patients diagnosed in 1990-1999 and 2000-2015?

No adjustments.

```
library(dplyr)
## Select, filter, and mutate necessary variables
surv_earlylate <- surv_exp %>%
  select(death, followup, dx_year) %>%
  filter(dx_year >= 1990) %>%
  mutate(dx_year = ifelse(dx_year >= 2000, yes = "late", no = "early"))
## Kaplan-Meier fit
library(survival)
surv_earlylate_fit <- with(surv_earlylate,
                           survfit(formula = Surv(time=followup, event=death) ~ dx_year))
## Comparison
surv_earlylate_comp <- with(surv_earlylate,
                           survdiff(formula = Surv(time=followup, event=death) ~ dx_year))
```

4.3.2. Cleaning fit outputs

```
library(broom)
library(dplyr)
## Main analyses
surv_exp_tidy <- data.frame()
surv_fit_type <- c("All", "III", "Non-III", "ICCC3 main",
                   "III by transplantation", "Non-III by transplantation",
                   "II by sub-classes", "IV by stage", "All by stage")
for (i in 1:length(surv_fit_type)) {
tidy_fit <- tidy(surv_exp_fits[[i]])
which_analysis <- data.frame(fit = rep(surv_fit_type[i], nrow(tidy_fit)))
surv_exp_tidy <- bind_rows(surv_exp_tidy, cbind(which_analysis, tidy_fit))
}
surv_exp_tidy$strata[is.na(surv_exp_tidy$strata)] <- "strata=none"
rm(tidy_fit, which_analysis)

## Early (1990-99) vs.late (2000-15) diagnosed comparison
surv_earlylate_fit  <- tidy(surv_earlylate_fit)
```

4.3.3. Tabulation

```
library(gridExtra)
library(dplyr)
## Fit summary columns:
# fit, strata, median follow-up, follow-up sum
```

```r
table_surv1 <- surv_exp_tidy %>%
  group_by(fit, strata) %>%
  summarise(min_time = min(time),
            max_time = max(time))
# Year-estimates and their CIs need to be collected
# from raw fit outputs
## 1. Empty matrix for collection
table_surv2 <- matrix(nrow = 30, ncol = 5)
### Columns: 1-year (CI), 5-year (CI), 10-year (CI), 15-year (CI), 20-year (CI)
## 2. For each fit and strata,
## 3. For each year-cutoff, grab the three needed values with stepfun, and
## paste them together in format "survival (lowerCI - upperCI)", and fill in
for (i in 1:nrow(table_surv1)) {
  df <- filter(surv_exp_tidy, fit == table_surv1$fit[i] & strata == table_surv1$strata[i])
  surv_est <- stepfun(df$time, c(1, df$estimate))
  surv_low <- stepfun(df$time, c(1, df$conf.low))
  surv_high <- stepfun(df$time, c(1, df$conf.high))
  year_est <- c(1, 5, 10, 15, 20)
  for (j in 1:length(year_est)) {
    if (year_est[j] <= max(df$time)) {
      table_surv2[i, j] <-
        paste(round(surv_est(year_est[j]), 2),
            " (",
            round(surv_low(year_est[j]), 2),
            " - ",
            round(surv_high(year_est[j]), 2),
            ")",
            sep = "")
      } else {
        table_surv2[i, j] <- NA
      }
  }
}
table_surv2 <- as.data.frame(table_surv2)
names(table_surv2) <- c("one_year_95CI",
                        "five_year_95CI",
                        "ten_year_95CI",
                        "fifteen_year_95CI",
                        "twenty_year_95CI")
table_surv <- bind_cols(table_surv1, table_surv2)
rm(table_surv1, table_surv2, i, j, surv_est, surv_high, surv_low, year_est, df)

## Comparison summary columns:
# Columns: group, compare, chisq, p_value
table_surv_comp <- data.frame(group = c("IV", "All", "All"),
                              compare = c("stage", "stage", "90's vs. 00's diagnosed"),
                              chisq = c(surv_exp_comp[[1]]$chisq,
                                        surv_exp_comp[[2]]$chisq,
                                        surv_earlylate_comp$chisq),
                              p_value = c(0.0366, 4.09e-12, 0.962))
# Plot tables
library(gridExtra)
grid.arrange(tableGrob(table_surv))
```

fit	strata	min_time	max_time	one_year_95CI	five_year_95CI	ten_year_95CI	fifteen_year_95CI	twenty_yea
All	strata=none	0.00000000	27.917864	0.95 (0.93 − 0.97)	0.84 (0.81 − 0.88)	0.83 (0.79 − 0.87)	0.82 (0.78 − 0.86)	0.82 (0.78 ·
All by stage	stage=1	0.64065708	27.167693	0.99 (0.97 − 1)	0.96 (0.94 − 0.99)	0.93 (0.89 − 0.97)	0.93 (0.89 − 0.97)	0.93 (0.89 ·
All by stage	stage=2	0.16153320	23.926078	0.97 (0.93 − 1)	0.95 (0.91 − 1)	0.94 (0.88 − 1)	0.94 (0.88 − 1)	0.94 (0.88
All by stage	stage=3	0.00000000	23.989049	0.92 (0.85 − 1)	0.82 (0.72 − 0.93)	0.82 (0.72 − 0.93)	0.82 (0.72 − 0.93)	0.82 (0.72 ·
All by stage	stage=4	0.03832991	27.917864	0.91 (0.85 − 0.98)	0.61 (0.5 − 0.73)	0.61 (0.5 − 0.73)	0.58 (0.47 − 0.71)	0.58 (0.47 ·
ICCC3 main	iccc_main=II	0.05749487	23.200548	0.95 (0.9 − 1)	0.89 (0.82 − 0.96)	0.89 (0.82 − 0.96)	0.89 (0.82 − 0.96)	0.89 (0.82 ·
ICCC3 main	iccc_main=III	0.00000000	23.926078	0.94 (0.9 − 0.98)	0.82 (0.76 − 0.89)	0.79 (0.72 − 0.86)	0.77 (0.69 − 0.85)	0.77 (0.69 ·
ICCC3 main	iccc_main=IV	1.06228611	21.409993	1 (1 − 1)	0.79 (0.66 − 0.94)	0.79 (0.66 − 0.94)	0.79 (0.66 − 0.94)	0.79 (0.66 ·
ICCC3 main	iccc_main=IX	0.23819302	24.939083	0.95 (0.88 − 1)	0.83 (0.71 − 0.96)	0.83 (0.71 − 0.96)	0.83 (0.71 − 0.96)	0.83 (0.71 ·
ICCC3 main	iccc_main=V	1.53319644	27.167693	1 (1 − 1)	0.92 (0.77 − 1)	0.92 (0.77 − 1)	0.92 (0.77 − 1)	0.92 (0.77
ICCC3 main	iccc_main=VI	0.11498973	27.917864	0.95 (0.87 − 1)	0.86 (0.75 − 0.98)	0.86 (0.75 − 0.98)	0.86 (0.75 − 0.98)	0.86 (0.75 ·
ICCC3 main	iccc_main=VII	0.12320329	8.432580	0.67 (0.38 − 1)	0.5 (0.22 − 1)	NA	NA	NA
ICCC3 main	iccc_main=VIII	0.25735797	18.439425	0.92 (0.77 − 1)	0.66 (0.43 − 1)	0.66 (0.43 − 1)	0.66 (0.43 − 1)	NA
ICCC3 main	iccc_main=X	1.72484600	19.783710	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	0.83 (0.58 − 1)	NA
ICCC3 main	iccc_main=XI	0.59958932	16.528405	1 (1 − 1)	0.86 (0.69 − 1)	0.86 (0.69 − 1)	0.86 (0.69 − 1)	NA
ICCC3 main	iccc_main=XII	2.14647502	11.753593	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	NA	NA
II by sub−classes	iccc=IIa	0.46543463	21.856263	0.97 (0.92 − 1)	0.97 (0.92 − 1)	0.97 (0.92 − 1)	0.97 (0.92 − 1)	0.97 (0.92
II by sub−classes	iccc=IIb	0.13415469	23.200548	0.93 (0.84 − 1)	0.78 (0.63 − 0.97)	0.78 (0.63 − 0.97)	0.78 (0.63 − 0.97)	0.78 (0.63 ·
II by sub−classes	iccc=IIc	0.05749487	16.678987	0.92 (0.79 − 1)	0.83 (0.64 − 1)	0.83 (0.64 − 1)	0.83 (0.64 − 1)	NA
II by sub−classes	iccc=IId	1.84804928	8.689938	1 (1 − 1)	1 (1 − 1)	NA	NA	NA
III	strata=none	0.00000000	23.926078	0.94 (0.9 − 0.98)	0.82 (0.76 − 0.89)	0.79 (0.72 − 0.86)	0.77 (0.69 − 0.85)	0.77 (0.69 ·
III by transplantation	hsct=0	0.00000000	23.926078	0.94 (0.9 − 0.98)	0.84 (0.78 − 0.9)	0.8 (0.73 − 0.87)	0.78 (0.7 − 0.86)	0.78 (0.7 −
III by transplantation	hsct=1	0.68993840	12.884326	1 (1 − 1)	0.4 (0.14 − 1)	0.4 (0.14 − 1)	NA	NA
IV by stage	stage=1	3.28542094	21.409993	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	1 (1 −
IV by stage	stage=2	2.26967830	14.587269	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	NA	NA
IV by stage	stage=3	20.47912389	20.479124	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	1 (1 − 1)	1 (1 −
IV by stage	stage=4	1.06228611	20.947296	1 (1 − 1)	0.57 (0.38 − 0.87)	0.57 (0.38 − 0.87)	0.57 (0.38 − 0.87)	0.57 (0.38 ·
Non−III	strata=none	0.05749487	27.917864	0.96 (0.93 − 0.98)	0.85 (0.8 − 0.89)	0.84 (0.79 − 0.89)	0.83 (0.78 − 0.88)	0.83 (0.78 ·
n−III by transplantation	hsct=0	0.05749487	27.917864	0.96 (0.93 − 0.99)	0.92 (0.88 − 0.95)	0.91 (0.87 − 0.95)	0.91 (0.87 − 0.95)	0.91 (0.87 ·
n−III by transplantation	hsct=1	0.76659822	16.366872	0.97 (0.93 − 1)	0.54 (0.4 − 0.73)	0.54 (0.4 − 0.73)	0.48 (0.33 − 0.7)	NA

```
grid.arrange(tableGrob(table_surv_comp))
```

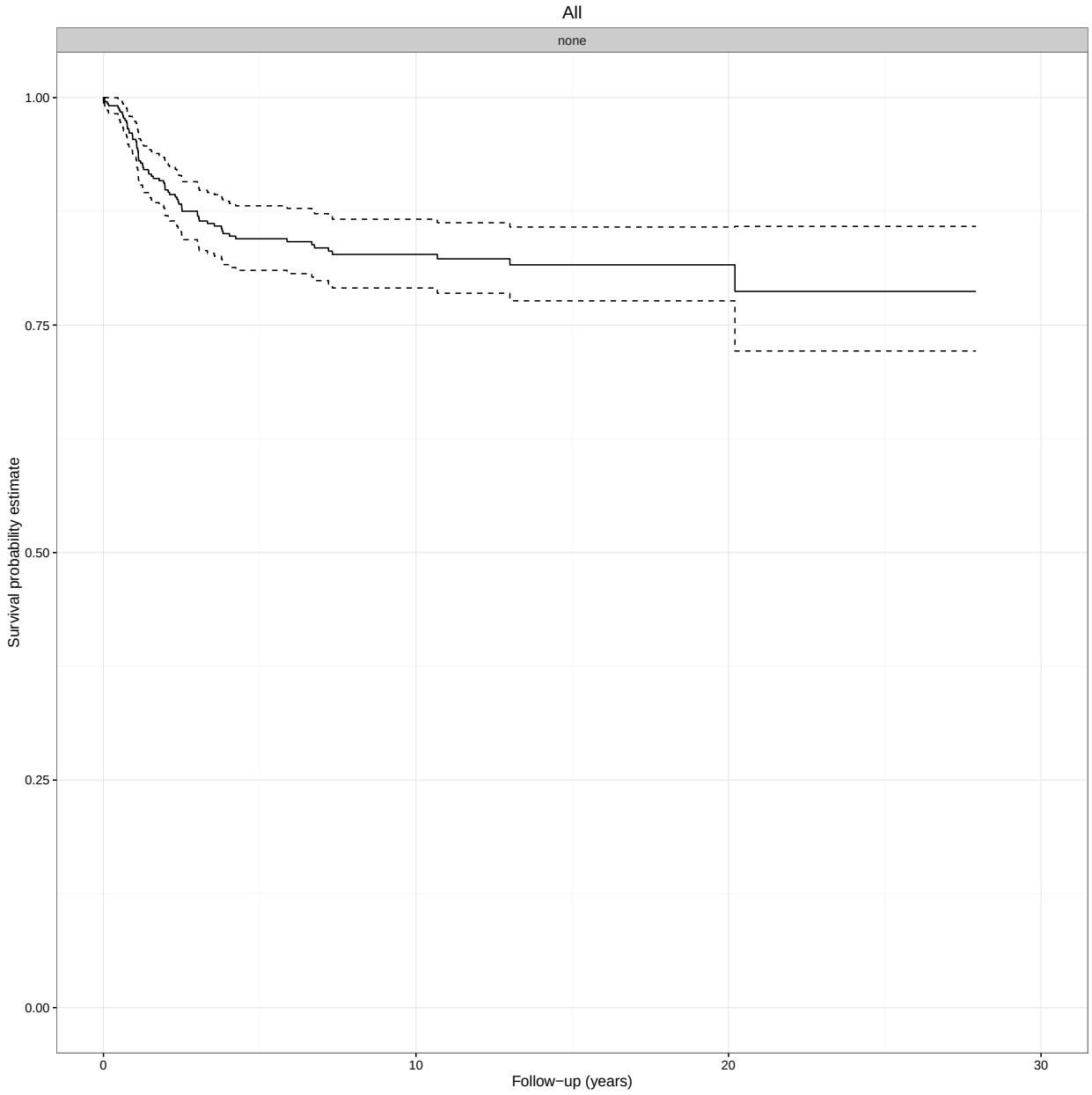
	group	compare	chisq	p_value
1	IV	stage	8.507570685	3.66e−02
2	All	stage	56.055489390	4.09e−12
3	All	90's vs. 00's diagnosed	0.002303577	9.62e−01

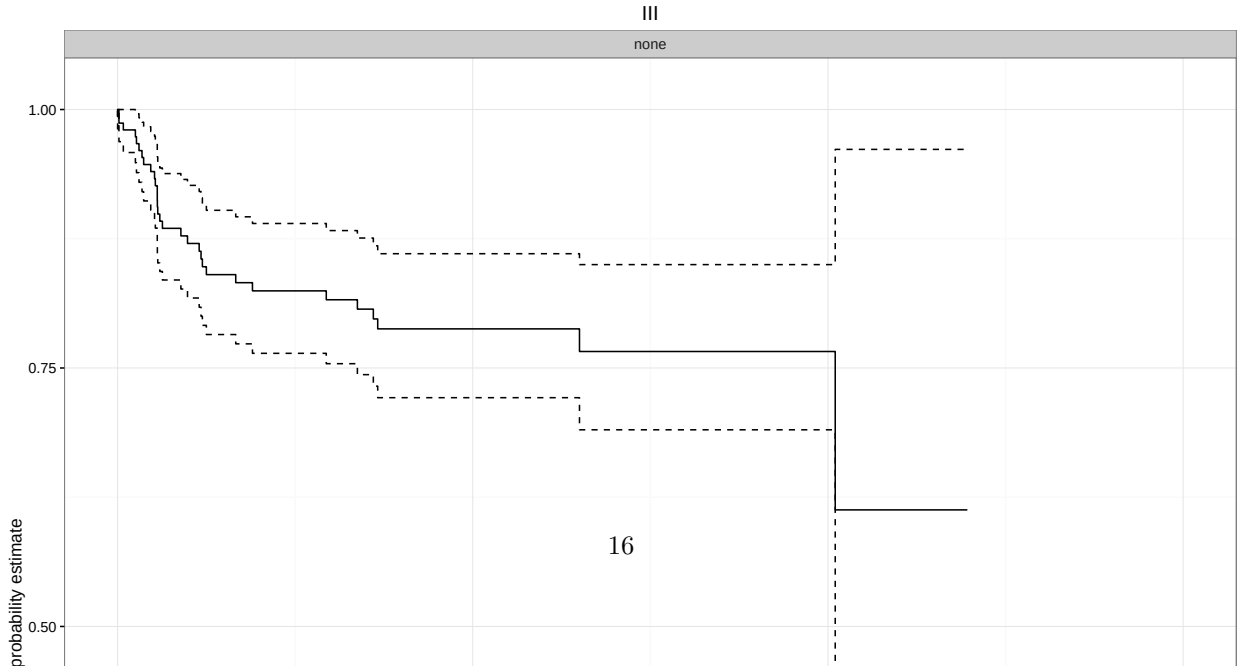
```
# setwd("figures&tables/protocol-output")
# pdf("survival-tables.pdf", width=14, height=12)
# grid.arrange(tableGrob(table_surv))
# grid.arrange(tableGrob(table_surv_comp))
# dev.off()
# for (i in 1:2) setwd("../")
```

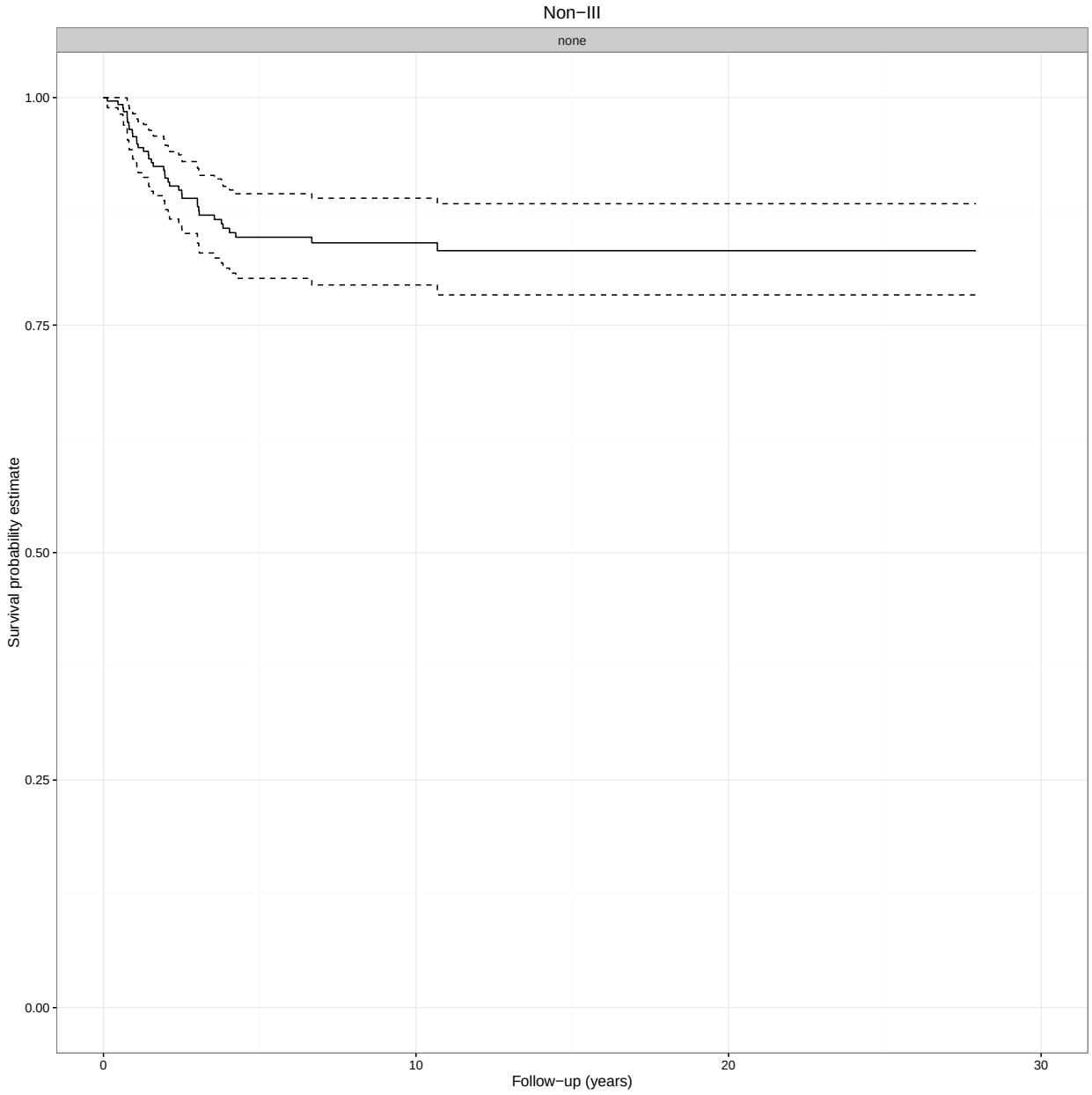
4.3.4. Visualization

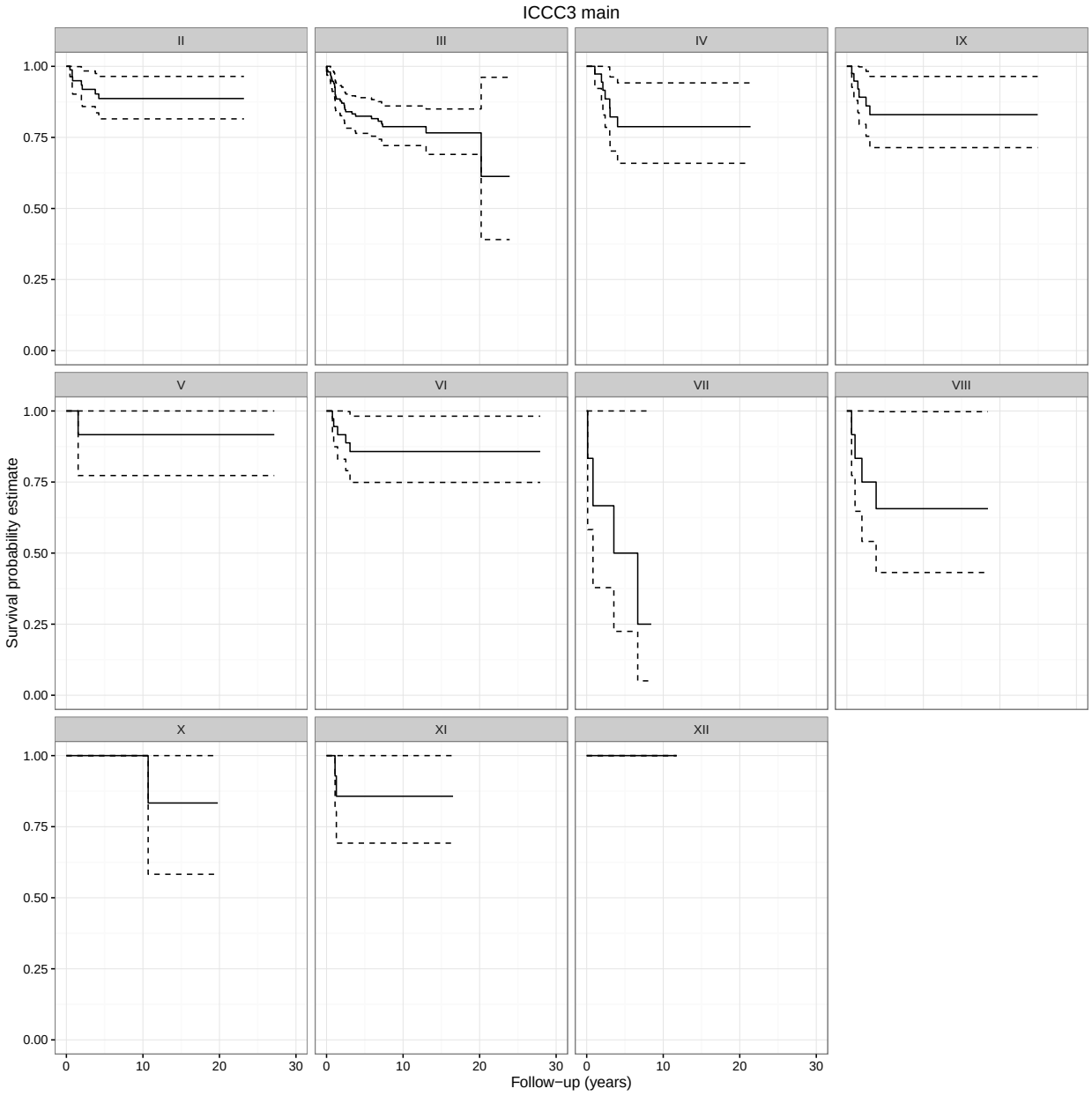
```
library(ggplot2)
library(tidyr)
library(dplyr)
```

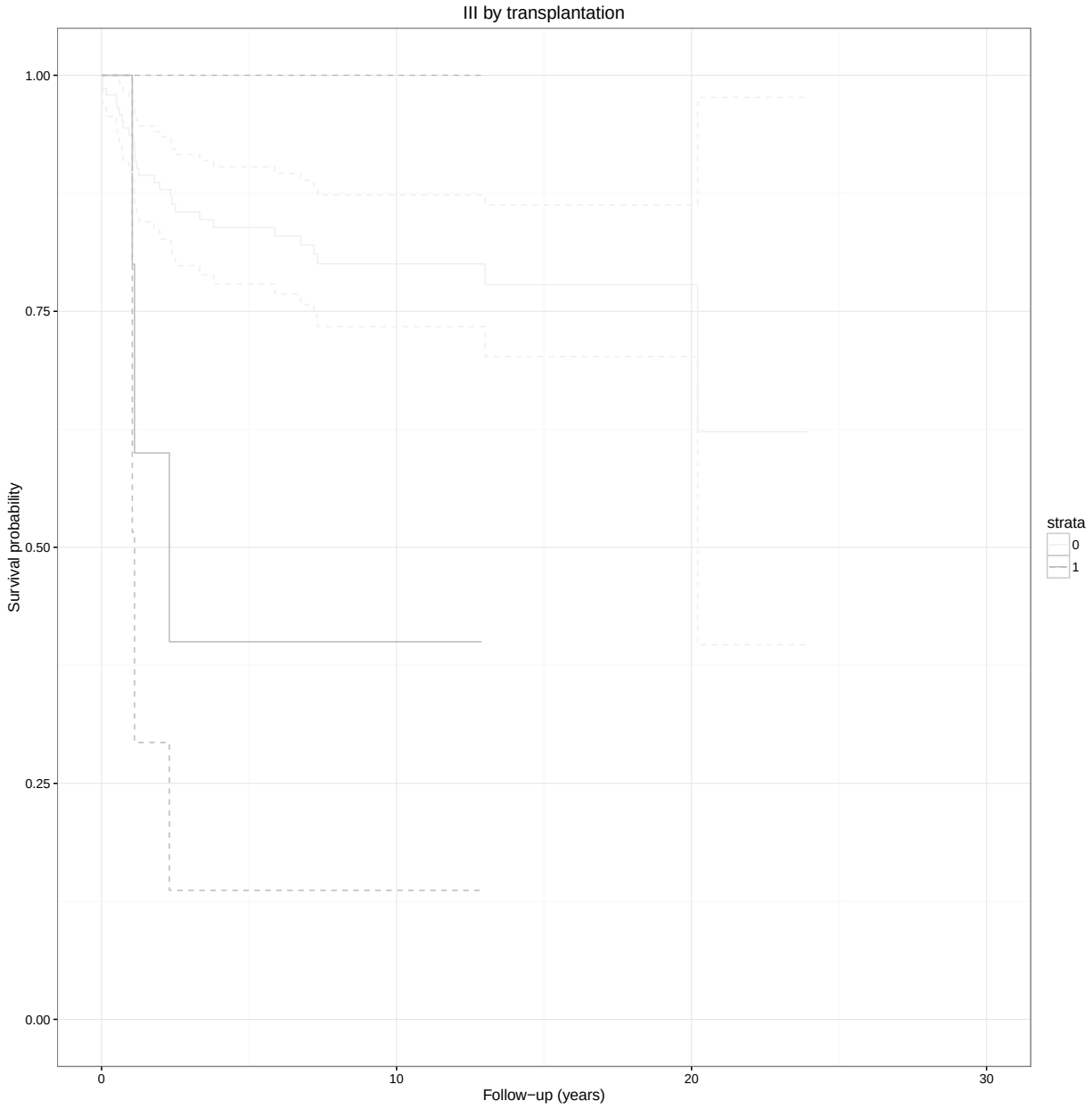
```r
# Plot, print and save all fits

## (x=0, y=1) points needs to be added for all fit-strata -pairs
zeros <- surv_exp_tidy %>%
  group_by(fit, strata) %>%
  summarise(time=0, estimate=1, conf.high=1, conf.low=1)
## Bind fits and (0, 1) points
surv_exp_tidy_plot <- bind_rows(surv_exp_tidy, zeros)
## Clean strata-variable
library(tidyr)
surv_exp_tidy_plot <- surv_exp_tidy_plot %>%
  separate(strata, c("delete", "strata"), sep="=") %>%
  select(-delete)

# Plotting
# setwd("figures&tables/protocol-output")
for (i in unique(surv_exp_tidy$fit)) {
  ## All, III, Non-III, ICCC3 main, and II are faceted
  if (i %in% c("All", "III", "Non-III", "ICCC3 main", "II by sub-classes")) {
  surv_plot <- ggplot(filter(surv_exp_tidy_plot, fit == i)) +
    geom_step(aes(x=time, y=estimate)) +
    geom_step(aes(x=time, y=conf.high), linetype="dashed") +
    geom_step(aes(x=time, y=conf.low), linetype="dashed") +
    facet_wrap(~ strata) +
    theme_bw() +
    scale_y_continuous(limits=c(0, 1)) +
    scale_x_continuous(limits=c(0, 30)) +
    ggtitle(i) +
    xlab("Follow-up (years)") +
    ylab("Survival probability estimate")
  } else {
    ## Other fits to the same facet
      surv_plot <- ggplot(filter(surv_exp_tidy_plot, fit == i)) +
    geom_step(aes(x=time, y=estimate, colour=strata)) +
    theme_bw() +
    geom_step(aes(x=time, y=conf.high, colour=strata), linetype="dashed") +
    geom_step(aes(x=time, y=conf.low, colour=strata), linetype="dashed") +
    scale_color_brewer(palette="Greys") +
    scale_y_continuous(limits=c(0, 1)) +
    scale_x_continuous(limits=c(0, 30)) +
    ggtitle(i) +
    xlab("Follow-up (years)") +
    ylab("Survival probability")
  }
  ggsave(filename = paste("surv-plot-",
                          tolower(gsub(" ", "-", i)),
                          ".pdf",
                          sep=""),
         plot = surv_plot,
         width=8,
         height=10)
  print(surv_plot)
}
```

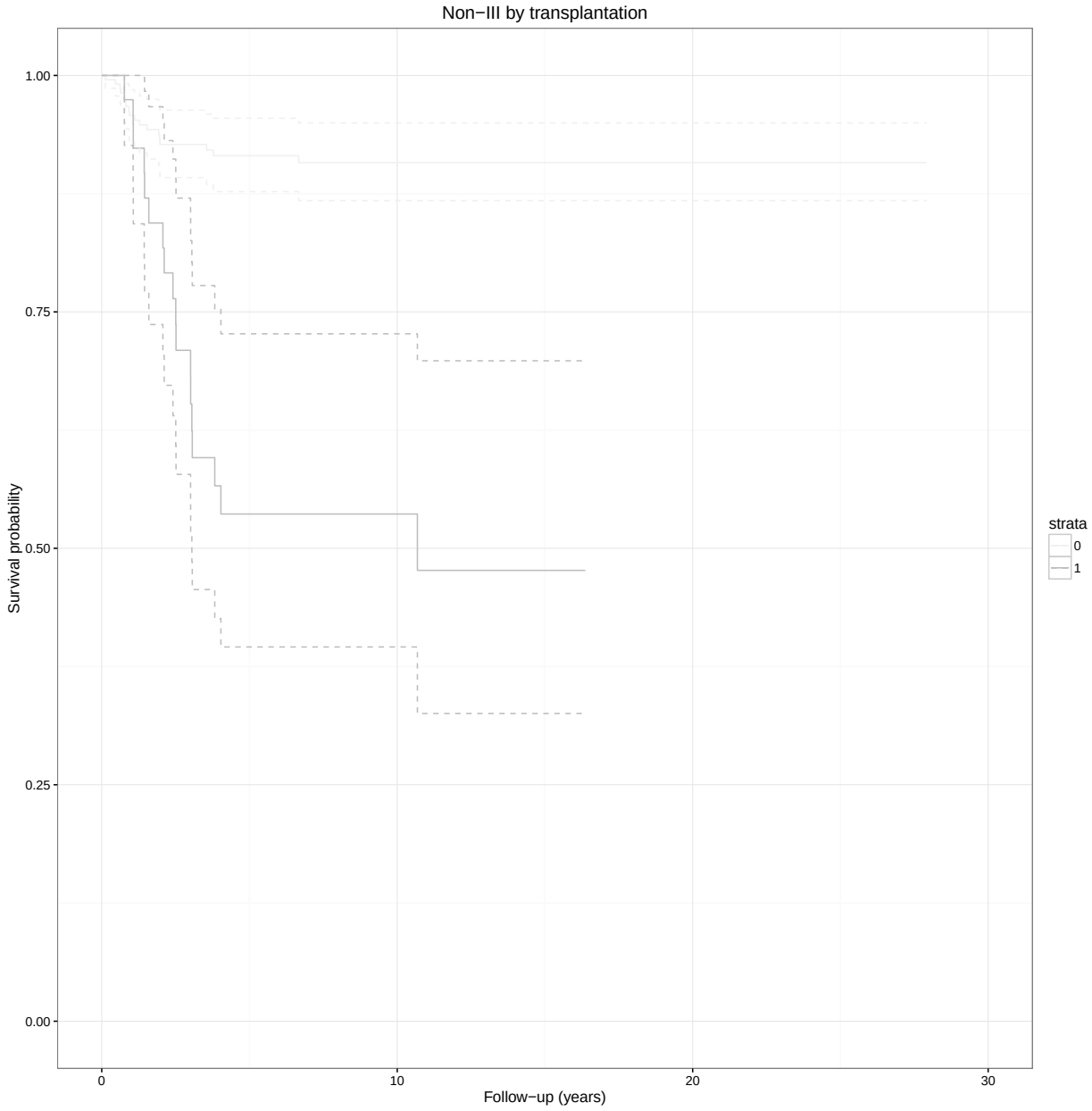

All
none
1.00
0.75
0.50
0.25
0.00
Survival probability estimate
0
10
20
30
Follow−up (years)

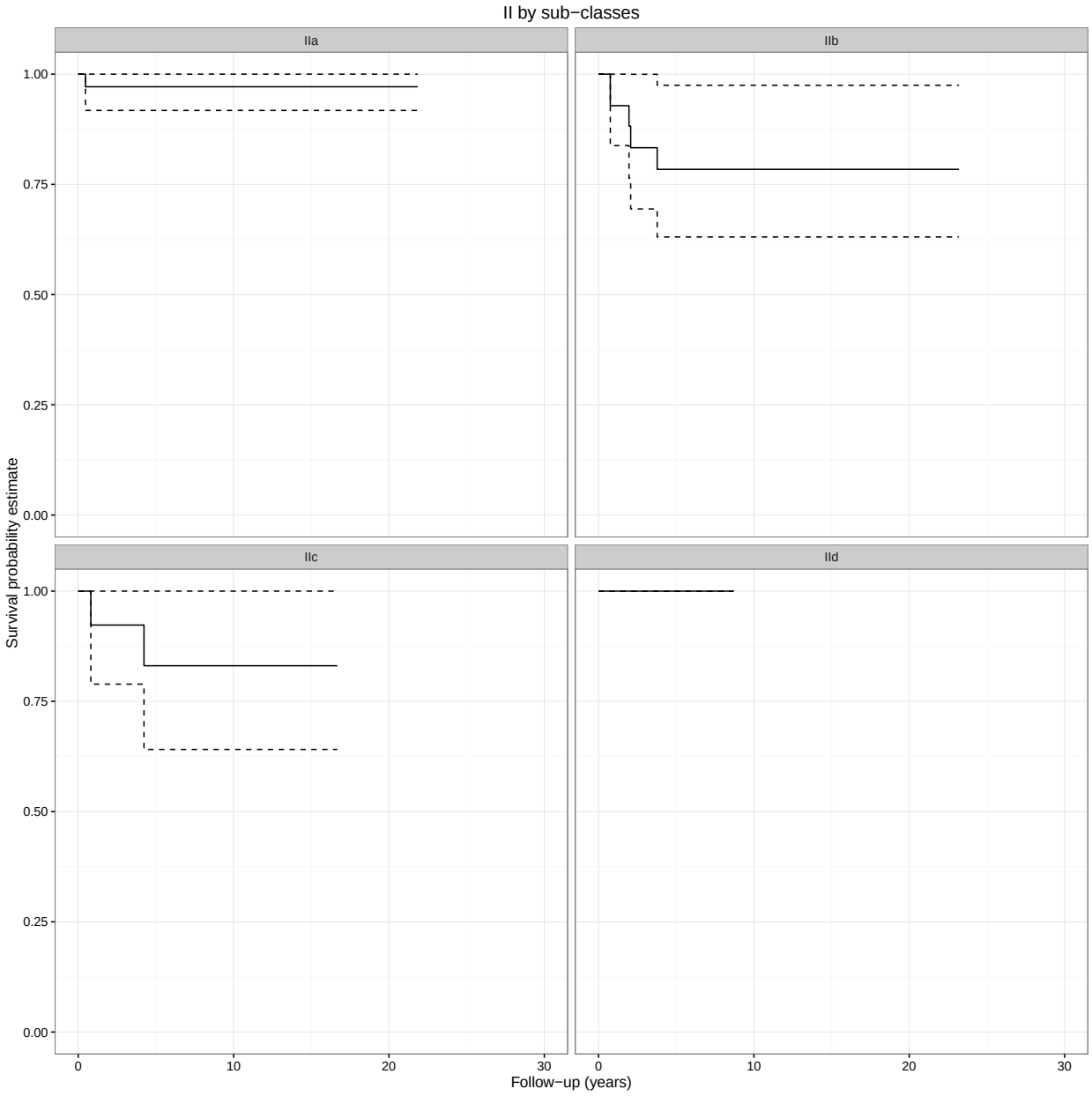

III
none
1.00
0.75
0.50
probability estimate

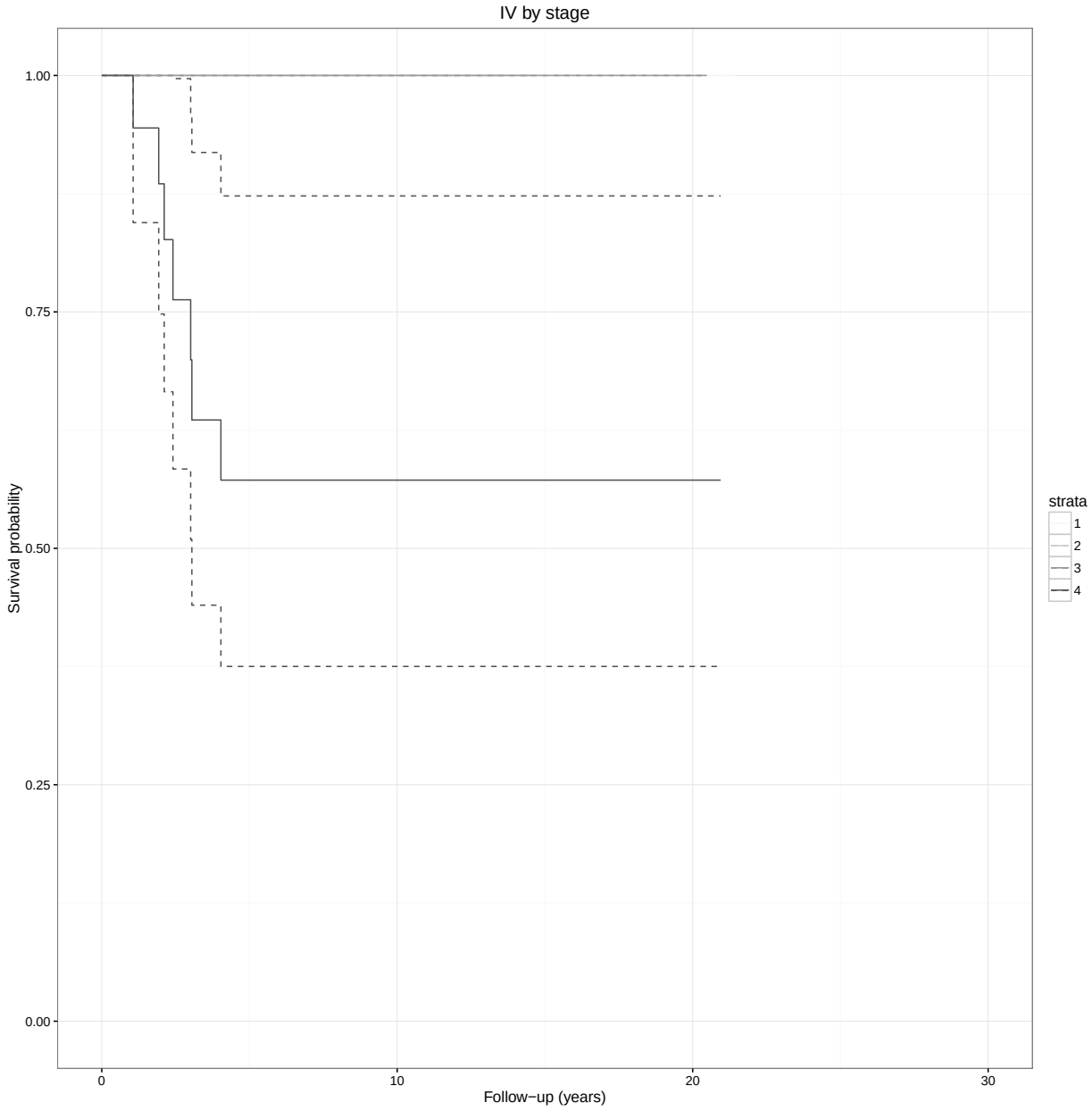

Non−III
none
1.00
0.75
0.50
0.25
0.00
Survival probability estimate
0
10
20
30
Follow−up (years)

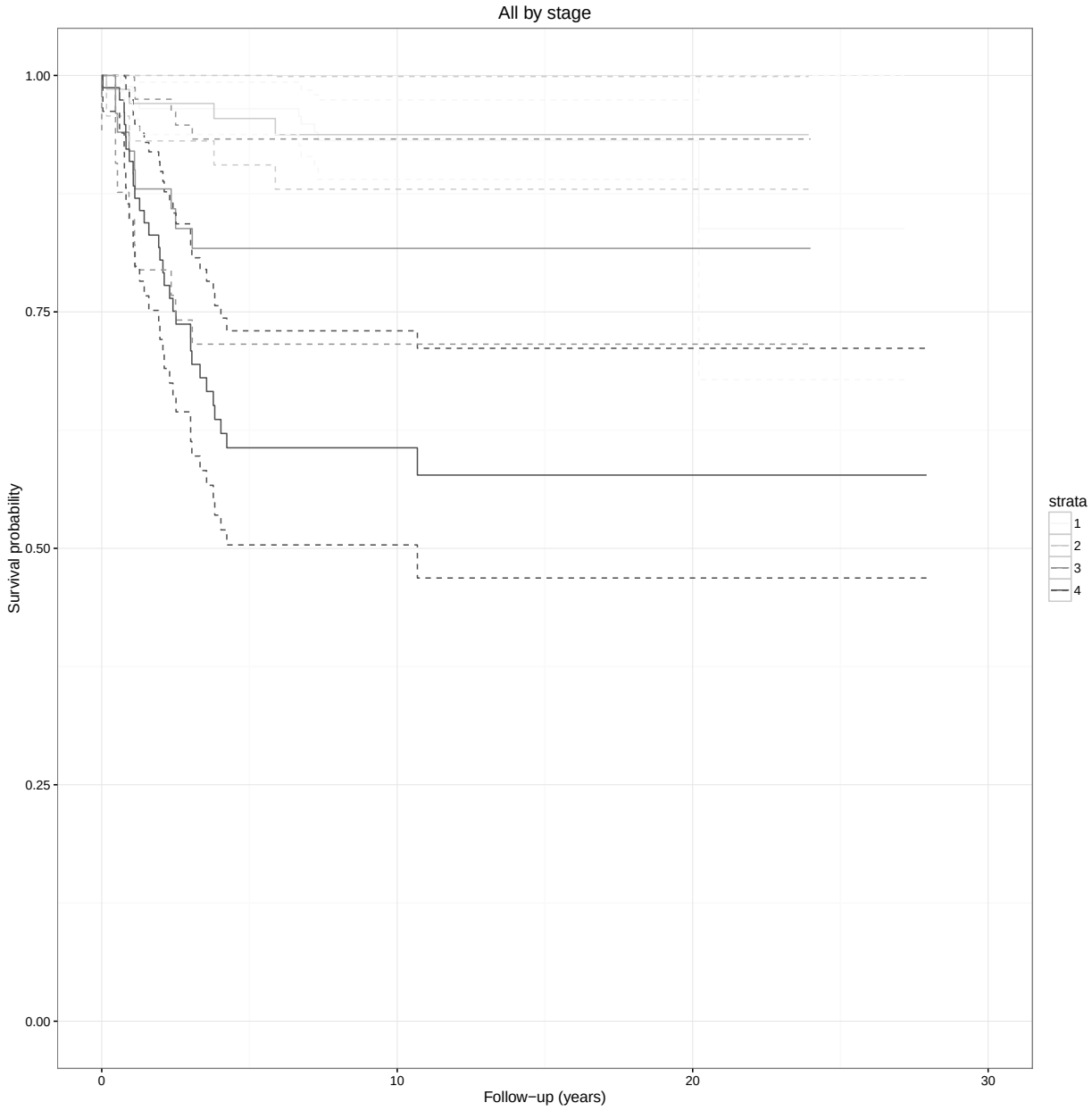

ICCC3 main
II
III
IV
IX
V
VI
VII
VIII
X
XI
XII
Survival probability estimate
1.00
0.75
0.50
0.25
0.00
0
10
20
30
Follow−up (years)


III by transplantation
1.00
0.75
0.50
0.25
0.00
Survival probability
0
10
20
30
Follow−up (years)
strata
0
1


Non−III by transplantation
1.00
0.75
0.50
0.25
0.00
Survival probability
0
10
20
30
Follow−up (years)
strata
0
1


II by sub−classes
IIa
IIb
IIc
IId
Survival probability estimate
1.00
0.75
0.50
0.25
0.00
0
10
20
30
Follow−up (years)


IV by stage
1.00
0.75
0.50
0.25
0.00
Survival probability
0
10
20
30
Follow−up (years)
strata
1
2
3
4

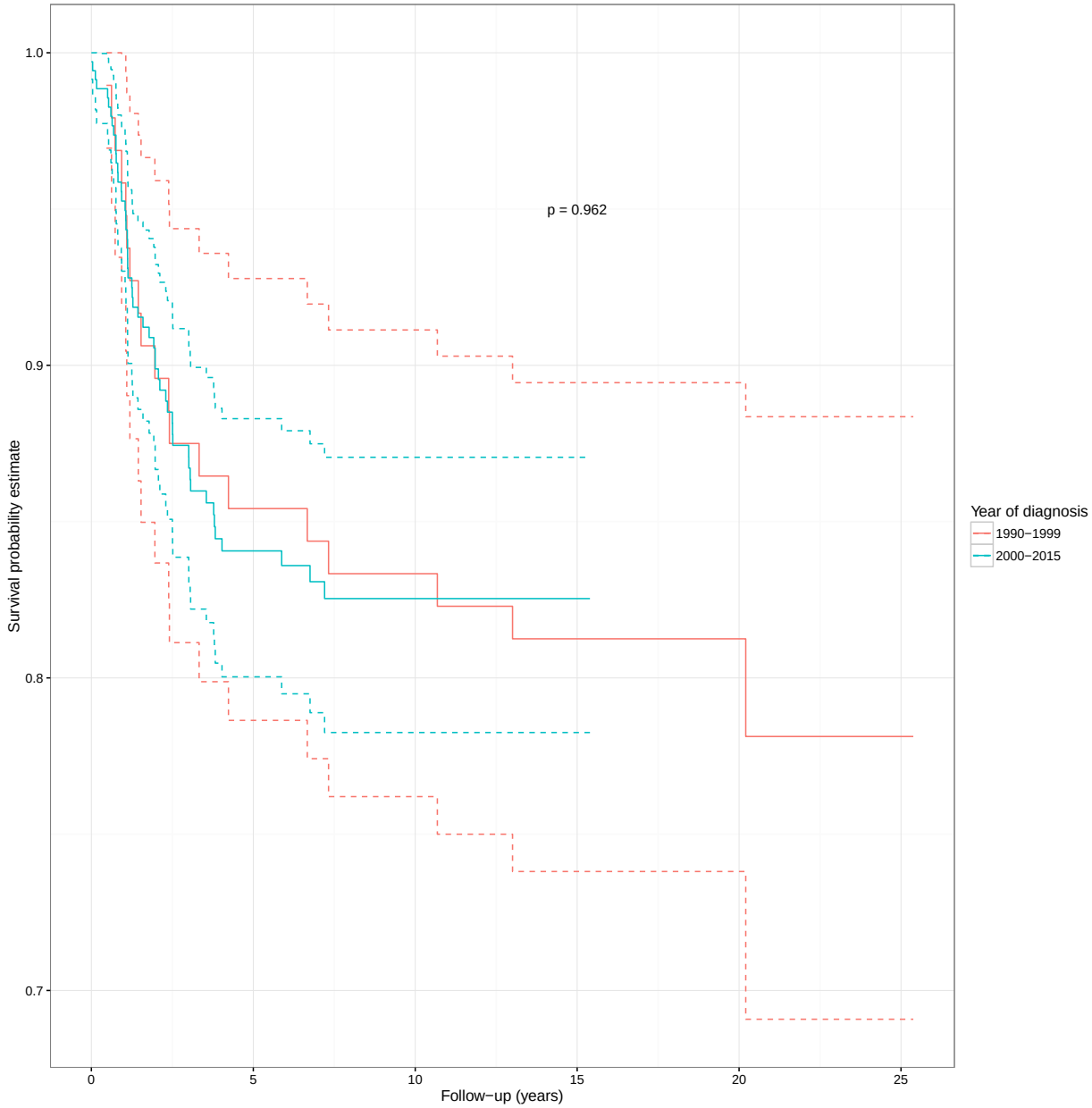



```
# for (i in 1:2) setwd("../")
rm(zeros)
```

4.4. How different was survival in patients diagnosed in 1990-1999 and 2000-2015?

```
library(ggplot2)
surv_earlylate_plot <- ggplot(surv_earlylate_fit) +
  geom_step(aes(x = time, y = estimate, color = strata)) +
  geom_step(aes(x = time, y = conf.high, color = strata),
            linetype = "dashed") +
```

```
  geom_step(aes(x = time, y = conf.low, color = strata),
            linetype = "dashed") +
  theme_bw() +
  xlab("Follow-up (years)") +
  ylab("Survival probability estimate") +
  scale_color_discrete(labels=c("1990-1999", "2000-2015")) +
  labs(color = "Year of diagnosis") +
  annotate("text",
           y=0.95,
           x=15,
           label=paste("p =", table_surv_comp$p_value[3]))

surv_earlylate_plot
```



```
# setwd("figures&tables/protocol-output")
# ggsave("surv-early-vs-late-diagnosed.pdf")
# for (i in 1:2) setwd("../")
```

4.5. How these results compare to European numbers and Finnish numbers?

- Europe 2000-2007: Gatta et al. PMID: 24314616
- Finland 2001-2010: Madanat-Harjuoja et al. PMID: 24623568

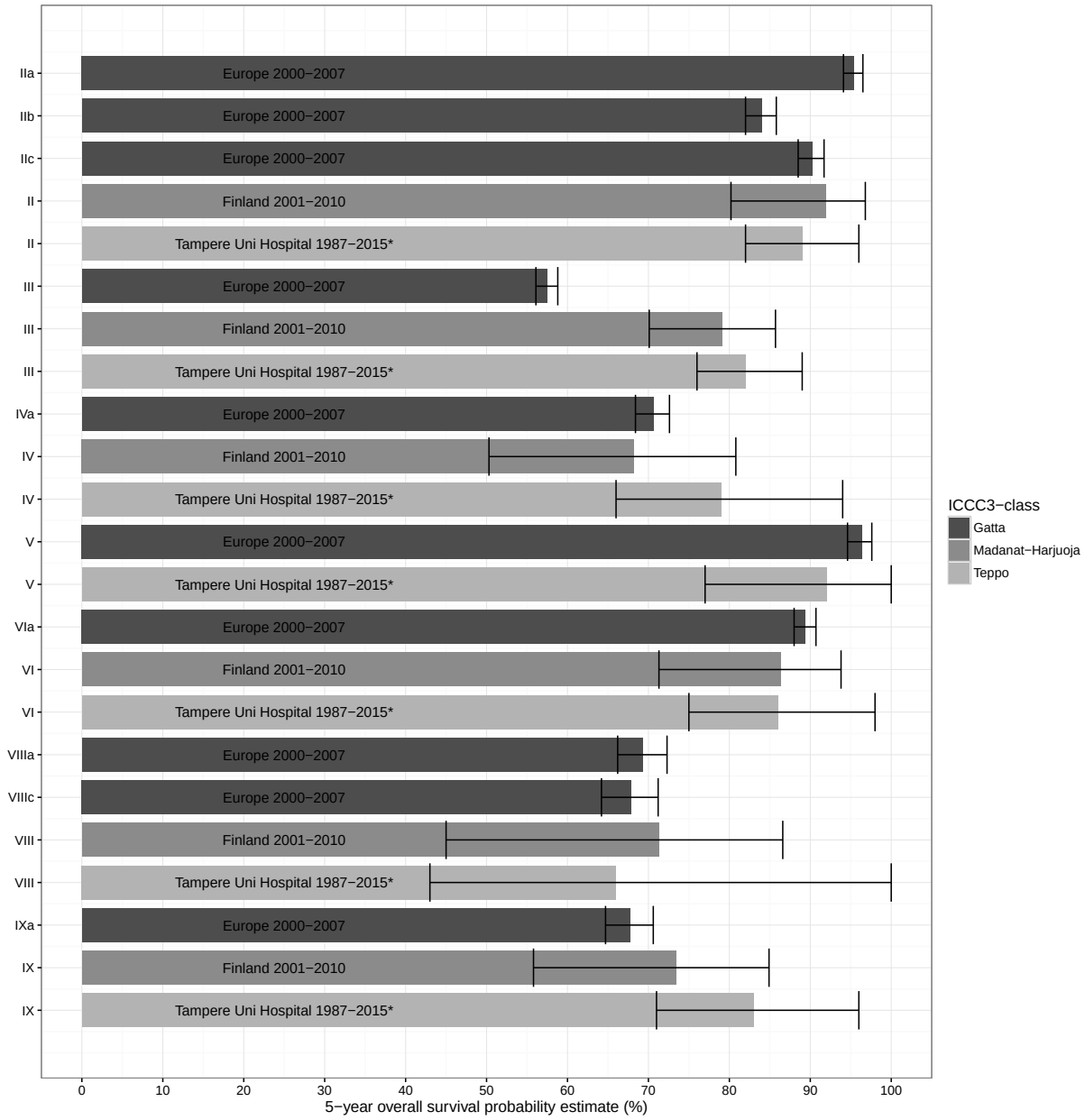
```
## Collecting 5 year overall survival from the articles
## Gatta et al. and Madanat-Harjuoja et al.
## and the results of the current study (Teppo et al.)
gatta_madanat <-
  data.frame(row = 1:23,
             iccc_main = c(rep("II", 3), "II", "II",
                           "III", "III", "III",
                           "IV", "IV", "IV",
                           "V", "V",
                           "VI", "VI", "VI",
                           "VIII", "VIII", "VIII", "VIII",
                           "IX", "IX", "IX"),
             iccc_sub = c("IIa", "IIb", "IIc", "II", "II",
                          "III", "III", "III",
                          "IVa", "IV", "IV",
                          "V", "V",
                          "VIa", "VI", "VI",
                          "VIIIa", "VIIIc", "VIII", "VIII",
                          "IXa", "IX", "IX"),
             dx_years=c(rep("2000-2007", 3),"2001-2010","1987-2015*",
                        "2000-2007", "2001-2010", "1987-2015*",
                        "2000-2007", "2001-2010", "1987-2015*",
                        "2000-2007", "1987-2015*",
                        "2000-2007", "2001-2010", "1987-2015*",
                        rep("2000-2007",2),"2001-2010", "1987-2015*",
                        "2000-2007", "2001-2010", "1987-2015*"),
             study = c(rep("Gatta", 3),"Madanat-Harjuoja","Teppo",
                       "Gatta", "Madanat-Harjuoja", "Teppo",
                       "Gatta", "Madanat-Harjuoja", "Teppo",
                       "Gatta", "Teppo",
                       "Gatta", "Madanat-Harjuoja", "Teppo",
                       rep("Gatta",2), "Madanat-Harjuoja", "Teppo",
                       "Gatta", "Madanat-Harjuoja", "Teppo"),
             area = c(rep("Europe", 3),"Finland","Tampere Uni Hospital",
                       "Europe", "Finland", "Tampere Uni Hospital",
                       "Europe", "Finland", "Tampere Uni Hospital",
                       "Europe", "Tampere Uni Hospital",
                       "Europe", "Finland", "Tampere Uni Hospital",
                       rep("Europe", 2),"Finland","Tampere Uni Hospital",
                       "Europe", "Finland", "Tampere Uni Hospital"),
             five_year_os = c(95.4, 84.0, 90.2, 91.9, 89,
                              57.5, 79.1, 82,
                              70.6, 68.2, 79,
```

```
                              96.4, 92,
                              89.4, 86.3, 86,
                              69.3, 67.9, 71.3, 66,
                              67.7, 73.4, 83),
             conf.low = c(94.1, 82.0, 88.5, 80.2, 82,
                          56.1, 70.1, 76,
                          68.4, 50.3, 66,
                          94.6, 77,
                          88, 71.3, 75,
                          66.2, 64.2, 45.0, 43,
                          64.7, 55.8, 71),
             conf.high = c(96.5, 85.8, 91.7, 96.8, 96,
                           58.8, 85.7, 89,
                           72.6, 80.8, 94,
                           97.6, 100,
                           90.7, 93.8, 98,
                           72.3, 71.2, 86.6, 100,
                           70.6, 84.9, 96))
gatta_madanat$iccc_main <- factor(gatta_madanat$iccc_main,
                                  levels=c("II", "III", "IV", "V", "VI", "VIII", "IX"))

library(gridExtra)
grid.arrange(tableGrob(gatta_madanat))
```

	row	iccc_main	iccc_sub	dx_years	study	area	five_year_os	conf.low	conf.high
1	1	II	IIa	2000–2007	Gatta	Europe	95.4	94.1	96.5
2	2	II	IIb	2000–2007	Gatta	Europe	84.0	82.0	85.8
3	3	II	IIc	2000–2007	Gatta	Europe	90.2	88.5	91.7
4	4	II	II	2001–2010	Madanat–Harjuoja	Finland	91.9	80.2	96.8
5	5	II	II	1987–2015*	Teppo	Tampere Uni Hospital	89.0	82.0	96.0
6	6	III	III	2000–2007	Gatta	Europe	57.5	56.1	58.8
7	7	III	III	2001–2010	Madanat–Harjuoja	Finland	79.1	70.1	85.7
8	8	III	III	1987–2015*	Teppo	Tampere Uni Hospital	82.0	76.0	89.0
9	9	IV	IVa	2000–2007	Gatta	Europe	70.6	68.4	72.6
10	10	IV	IV	2001–2010	Madanat–Harjuoja	Finland	68.2	50.3	80.8
11	11	IV	IV	1987–2015*	Teppo	Tampere Uni Hospital	79.0	66.0	94.0
12	12	V	V	2000–2007	Gatta	Europe	96.4	94.6	97.6
13	13	V	V	1987–2015*	Teppo	Tampere Uni Hospital	92.0	77.0	100.0
14	14	VI	VIa	2000–2007	Gatta	Europe	89.4	88.0	90.7
15	15	VI	VI	2001–2010	Madanat–Harjuoja	Finland	86.3	71.3	93.8
16	16	VI	VI	1987–2015*	Teppo	Tampere Uni Hospital	86.0	75.0	98.0
17	17	VIII	VIIIa	2000–2007	Gatta	Europe	69.3	66.2	72.3
18	18	VIII	VIIIc	2000–2007	Gatta	Europe	67.9	64.2	71.2
19	19	VIII	VIII	2001–2010	Madanat–Harjuoja	Finland	71.3	45.0	86.6
20	20	VIII	VIII	1987–2015*	Teppo	Tampere Uni Hospital	66.0	43.0	100.0
21	21	IX	IXa	2000–2007	Gatta	Europe	67.7	64.7	70.6
22	22	IX	IX	2001–2010	Madanat–Harjuoja	Finland	73.4	55.8	84.9
23	23	IX	IX	1987–2015*	Teppo	Tampere Uni Hospital	83.0	71.0	96.0

```
## Flipped bargraph; bar by row (fine-tuning done later by hand)
library(ggplot2)
gatta_madanat_plot <- ggplot(gatta_madanat) +
  geom_bar(aes(x = row, y = five_year_os, fill = study, width=0.8), stat="identity") +
  geom_errorbar(aes(x = row, ymin = conf.low, ymax = conf.high)) +
  coord_flip() +
  theme_bw() +
  scale_x_continuous(breaks=1:23, labels=gatta_madanat$iccc_sub, trans="reverse") +
  scale_y_continuous(breaks=seq(0, 100, by=10)) +
  geom_text(aes(x=row, y=25, label=paste(area, dx_years, sep=" ")), size = 4) +
  ylab("5-year overall survival probability estimate (%)") +
  xlab("") +
  labs(fill="ICCC3-class") +
  scale_fill_grey(start=0.3, end=0.7)
```

```
gatta_madanat_plot
```




```
# setwd("figures&tables/protocol-output")
# pdf("gatta-madanat-table.pdf", width=14, height=8)
# grid.arrange(tableGrob(gatta_madanat))
# dev.off()
ggsave("survival-gatta-madanat.pdf")
# for (i in 1:2) setwd("../")
```

4.6. How different was survival during the first 5 years and after surviving 5 years?

This analysis is about the concentration of the risk of death to the time after diagnosis. Threshold of 5 years is used because it's a reasonable convention and the main results are reported with this threshold.

5. Batch 2: The stage distribution

5.1. Selecting from clean data

```
library(dplyr)
stage <- pottikk %>%
  select(iccc_main, stage)
```

5.2. Describing and Handling missing values

Missingness:

- ICCC3-class missing: 30 out of 447
- Stage missing: 80 out of 447
- Both missing: 102 out of 447

Handling: Discarding incomplete observations (447 cases -> 345 cases)

```
stage <- stage[complete.cases(stage), ]
```

5.3. Analysis

- Primary analytics method: Simultaneous confidence intervals for multinomial proportions

```
# install.packages("DescTools")
library(DescTools)
# Empty list for results
stage_ci <- data.frame()
for (i in 1:length(unique(stage$iccc_main))) {
  sb <- stage[stage$iccc_main == unique(stage$iccc_main)[i], ]
  occur <- c(table(factor(sb$stage, levels=as.character(1:4))))
  complete <- cbind(iccc_main = rep(unique(stage$iccc_main)[i], 4),
                    stage = as.character(1:4),
                    n_per_iccc_main = rep(
                      nrow(stage[stage$iccc_main == unique(stage$iccc_main)[i], ]),
                      4),
                    round(as.data.frame(MultinomCI(occur)), 2))
  stage_ci <- rbind(stage_ci, complete)
}
rm(sb, occur, complete)
```

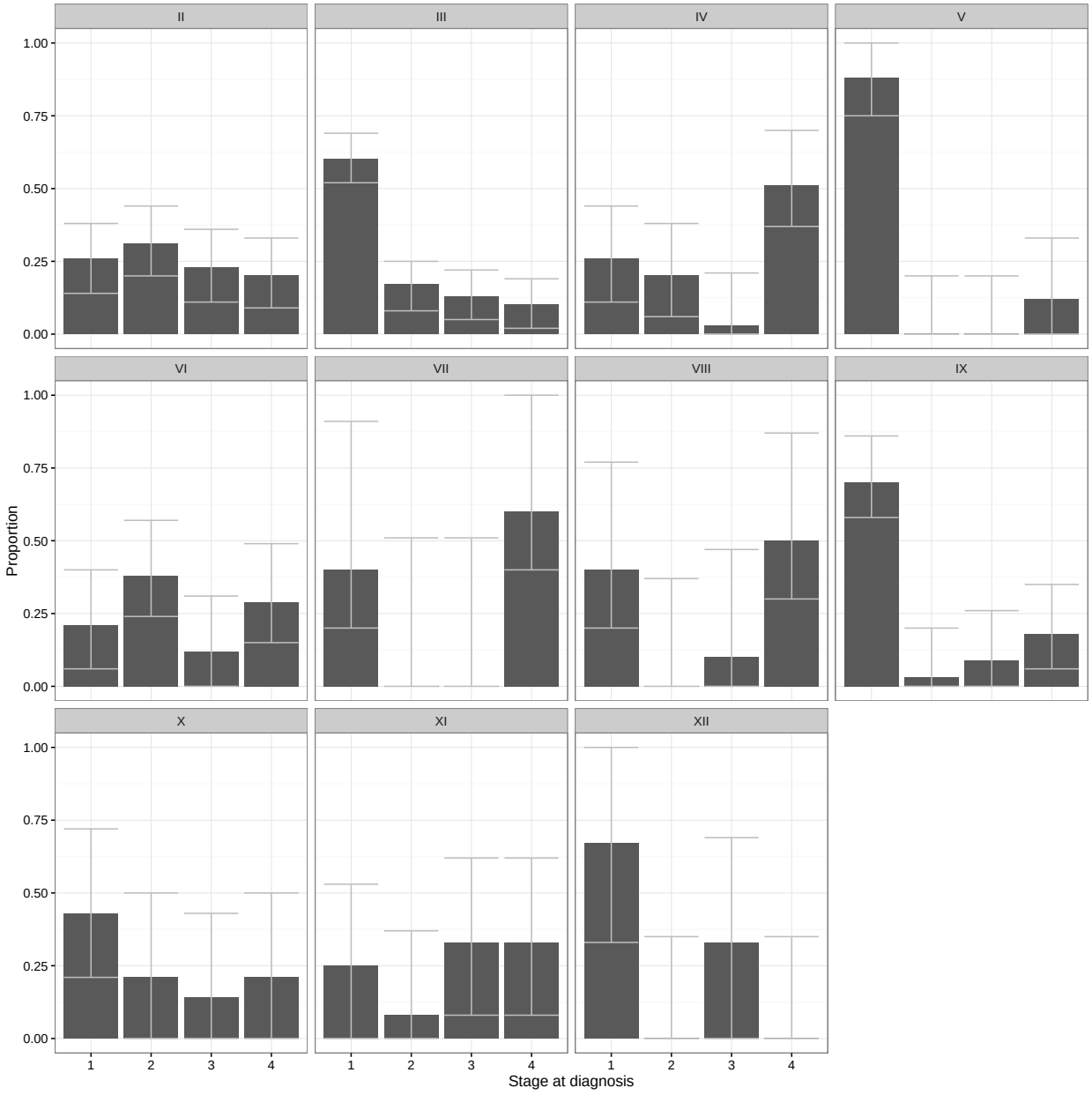
5.4. Tabulation

```
library(gridExtra)
library(ggplot2)
# setwd("figures&tables/protocol-output")
# pdf("stage-table.pdf", width=4, height=15)
# grid.arrange(tableGrob(stage_ci))
# dev.off()
# for (i in 1:2) setwd("../")
grid.arrange(tableGrob(stage_ci))
```

1	II	1	70	0.26	0.14	0.38
2	II	2	70	0.31	0.20	0.44
3	II	3	70	0.23	0.11	0.36
4	II	4	70	0.20	0.09	0.33
11	IV	1	35	0.26	0.11	0.44
21	IV	2	35	0.20	0.06	0.38
31	IV	3	35	0.03	0.00	0.21
41	IV	4	35	0.51	0.37	0.70
12	VI	1	34	0.21	0.06	0.40
22	VI	2	34	0.38	0.24	0.57
32	VI	3	34	0.12	0.00	0.31
42	VI	4	34	0.29	0.15	0.49
13	VIII	1	10	0.40	0.20	0.77
23	VIII	2	10	0.00	0.00	0.37
33	VIII	3	10	0.10	0.00	0.47
43	VIII	4	10	0.50	0.30	0.87
14	IX	1	33	0.70	0.58	0.86
24	IX	2	33	0.03	0.00	0.20
34	IX	3	33	0.09	0.00	0.26
44	IX	4	33	0.18	0.06	0.35
15	X	1	14	0.43	0.21	0.72
25	X	2	14	0.21	0.00	0.50
35	X	3	14	0.14	0.00	0.43
45	X	4	14	0.21	0.00	0.50
16	III	1	121	0.60	0.52	0.69
26	III	2	121	0.17	0.08	0.25
36	III	3	121	0.13	0.05	0.22
46	III	4	121	0.10	0.02	0.19
17	XI	1	12	0.25	0.00	0.53
27	XI	2	12	0.08	0.00	0.37
37	XI	3	12	0.33	0.08	0.62
47	XI	4	12	0.33	0.08	0.62
18	VII	1	5	0.40	0.20	0.91
28	VII	2	5	0.00	0.00	0.51
38	VII	3	5	0.00	0.00	0.51
48	VII	4	5	0.60	0.40	1.00
19	V	1	8	0.88	0.75	1.00
29	V	2	8	0.00	0.00	0.20
39	V	3	8	0.00	0.00	0.20
49	V	4	8	0.12	0.00	0.33
110	XII	1	3	0.67	0.33	1.00
210	XII	2	3	0.00	0.00	0.35
310	XII	3	3	0.33	0.00	0.69

5.5. Visualization

```
## Re-arrange ICCC-levels for plot
stage_plot <- stage_ci
stage_plot$iccc_main <- factor(stage_plot$iccc_main,
                               levels=c("II", "III", "IV", "V",
                                        "VI", "VII", "VIII",
                                        "IX", "X", "XI", "XII"))
## Plot
library(ggplot2)
ggplot(stage_plot, aes(x=stage)) +
  geom_bar(aes(y=est), stat="identity", position="stack") +
  facet_wrap(~ iccc_main) +
  theme_bw() +
  geom_errorbar(aes(x=stage, ymax=upr.ci, ymin=lwr.ci), color="grey") +
  ylab("Proportion") +
  xlab("Stage at diagnosis")
```




```
## Save
# setwd("figures&tables/protocol-output")
ggsave("stage-plot.pdf")
# for (i in 1:2) setwd("../")
```

6. Batch 3. Tumor size distribution comparison inference:

- Was there a difference between tumor size distributions of
  - a. patient populations with different stage at diagnosis?
  - b. III and non-III groups of patients?

- What was the difference between tumor size distributions of males and females in III and non-III patient populations?
- What was the correlation of tumor size and age at diagnosis in III and non-III patient populations?
- What was the correlation of tumor size and body area in non-CNS patients?

6.1. Selecting from clean data

```
library(dplyr)
size <- pottikk %>%
  select(sex, age_dx, iccc_cns, stage, tumorsize)
```

6.2. Describing and Handling missing values

Missingness structure:

- Tumor size has 136 missing values (Proportion: 30.4%)
- Tumor size and sex missing: 136
- Tumor size and age missing: 136
- Tumor size and stage missing: 181
- Tumor size and diagnosis (ICCC3) missing: 155

```
library(dplyr)
size <- pottikk %>%
  select(sex, age_dx, iccc_cns, stage, tumorsize) %>%
  filter(!is.na(tumorsize)) %>%
  arrange(tumorsize) %>%
  mutate(sex = factor(sex), iccc_cns = as.character(iccc_cns), stage = factor(stage))
```

Handling: Complete case analysis (discard incomplete cases) in all questions.

6.3. Analysis

6.3.1. Description

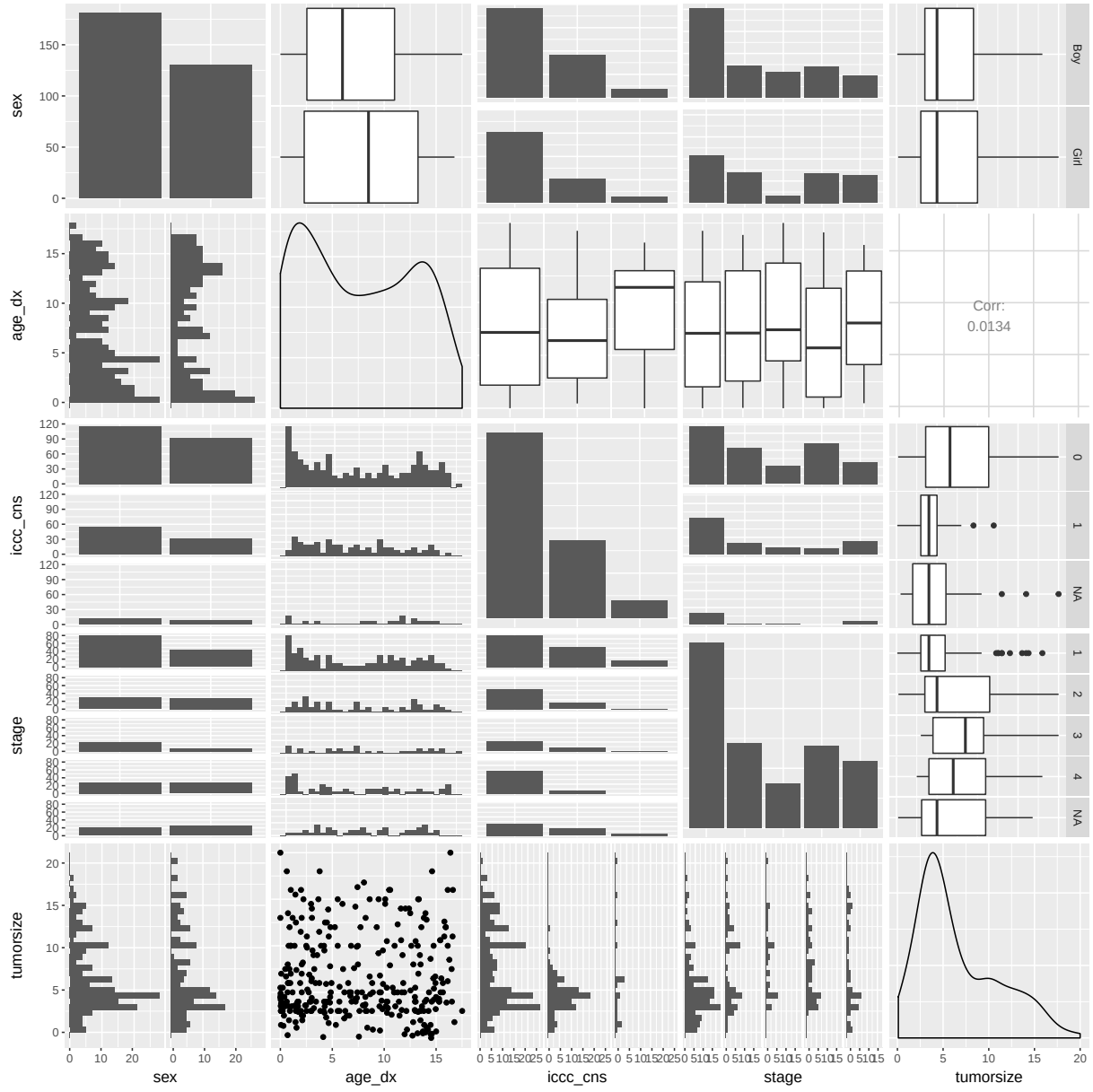
```
library(broom)
library(dplyr)
size_summary <- data.frame()
diagnosis <- data.frame(diagnosis = c("All", rep("All", 4), "III", "Non-III",
                                      rep("III", 2), rep("Non-III", 2)))
strata <-   data.frame(strata = c("None", paste("Stage", 1:4),
                                  rep("None", 2),  "Boy", "Girl",
                                  "Boy", "Girl"))
strata_index <- list(rep(TRUE, nrow(size)),
                     size$stage == "1",
                     size$stage == "2",
                     size$stage == "3",
                     size$stage == "4",
```

```
                      size$iccc_cns == "1",
                      size$iccc_cns == "0",
                      size$iccc_cns == "1" & size$sex == "Boy",
                      size$iccc_cns == "1" & size$sex == "Girl",
                      size$iccc_cns == "0" & size$sex == "Boy",
                      size$iccc_cns == "0" & size$sex == "Girl")
for (i in 1:nrow(strata)) {
  df <- tidy(summary(size$tumorsize[strata_index[[i]]]))[1:6]
  size_summary <- bind_rows(size_summary, df)
}
size_summary <- bind_cols(diagnosis, strata, size_summary)
rm(diagnosis, strata, strata_index, df)
```

6.3.2. Inference

- Checking variable and pairwise distributions

```
library(ggplot2)
library(GGally)
ggpairs(size)
```



Six non-parametric tests in total without multiple correction:

- Tumor size in stage groups = Kruskal-Wallis
- Tumor size in cns or non-cns groups = Mann-Whitney
- Tumor size in boys and girls (III and non-III groups) = Mann-Whitney
- Tumor size and age (III and non-III groups) = Pearson correlation coef. test

```
## Empty list for test results
size_tests <- list()
## Tests
size_tests[[1]] <- kruskal.test(tumorsize ~ stage,
                                data = size)
size_tests[[2]] <- wilcox.test(tumorsize ~ iccc_cns,
                               data = size,
```

```
                              conf.int=TRUE)
size_tests[[3]] <- wilcox.test(tumorsize ~ sex,
                              data = subset(size, iccc_cns == "1"),
                              conf.int=TRUE)
size_tests[[4]] <- wilcox.test(tumorsize ~ sex,
                              data = subset(size, iccc_cns == "0"),
                              conf.int=TRUE)
size_tests[[5]] <- cor.test(formula = ~ tumorsize + age_dx,
                            data = subset(size, iccc_cns == "1"))
size_tests[[6]] <- cor.test(formula = ~ tumorsize + age_dx,
                            data = subset(size, iccc_cns == "0"))
```

In addition, pair-wise comparison of tumor sizes in stage groups 1 - 4 is done with pair-wise Wilcoxon tests with Bonferroni correction.

```
with(size, pairwise.wilcox.test(tumorsize, stage, p.adjust.method = "bonferroni"))
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  tumorsize and stage
##
##   1       2       3
## 2 0.02940 -       -
## 3 0.00030 1.00000 -
## 4 0.00023 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

6.4. Cleaning analysis outputs

```
## Analyses
size_analyses <- data.frame(
in_group = c("All", "All", "III", "Non-III", "III", "Non-III"),
compare_by = c("Stage", "III/Non-III", "Sex", "Sex",
                    "Age at diagnosis", "Age at diagnosis"),
method = c("Kruskal-Wallis", rep("Mann-Whitney-Wilcoxon", 3),
                  rep("Pearson correlation coefficient", 2))
)
## Clean result-list
library(broom)
library(dplyr)

size_test_tidy <- data.frame()
for (i in 1:6) {
  tidy_test <- tidy(size_tests[[i]])
  size_test_tidy <- bind_rows(size_test_tidy, tidy_test)
}
size_tests <- bind_cols(size_analyses, size_test_tidy)
rm(size_test_tidy, tidy_test, size_analyses)
```

6.5. Tabulation

```
library(dplyr)
library(gridExtra)

## Description
grid.arrange(tableGrob(size_summary))
```

	diagnosis	strata	minimum	q1	median	mean	q3	maximum
1	All	None	0.10	3.30	5.0	6.572	10.00	20
2	All	Stage 1	0.10	3.00	4.0	5.055	6.00	18
3	All	Stage 2	0.24	3.50	5.0	7.438	11.50	20
4	All	Stage 3	3.00	4.50	8.5	8.463	10.75	20
5	All	Stage 4	2.50	4.00	7.0	7.780	11.00	18
6	III	None	0.10	3.00	4.0	4.252	5.00	12
7	Non−III	None	0.20	3.60	6.6	7.606	11.38	20
8	III	Boy	0.10	3.00	4.0	4.392	5.55	12
9	III	Girl	0.24	3.00	4.0	4.005	5.00	8
10	Non−III	Boy	0.35	3.85	6.5	7.530	10.60	18
11	Non−III	Girl	0.20	3.20	6.7	7.701	12.00	20

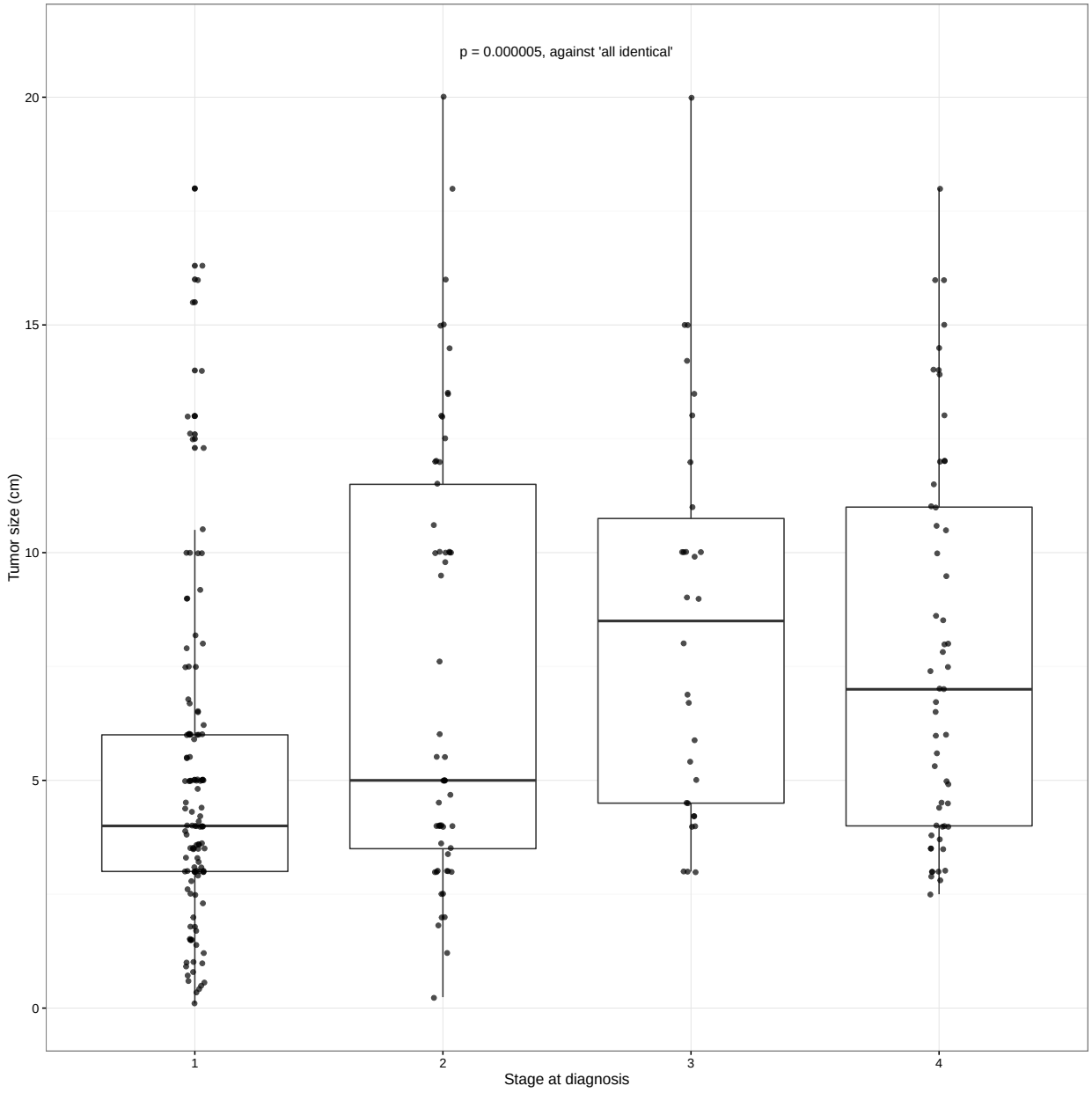
```
## Inference
size_tests <- size_tests %>% rename(df = parameter, test_stat = statistic)
grid.arrange(tableGrob(size_tests))
```

	in_group	compare_by	method	test_stat	p.value	df	estimate	conf.low	conf.high
1	All	Stage	Kruskal–Wallis	27.3149678	5.056810e−06	3	NA	NA	NA
2	All	III/Non−III	Mann−Whitney−Wilcoxon	12209.5000000	3.402094e−07	NA	2.500066e+00	1.40006696	3.9999917
3	III	Sex	Mann−Whitney−Wilcoxon	902.0000000	6.582446e−01	NA	4.938486e−05	−0.50005458	1.0000070
4	Non−III	Sex	Mann−Whitney−Wilcoxon	5231.0000000	9.981209e−01	NA	−7.135118e−05	−1.10004207	1.0000298
5	III	Age at diagnosis	Pearson correlation coefficient	−2.2652275	2.607222e−02	84	−2.399368e−01	−0.42995463	−0.0295644
6	Non−III	Age at diagnosis	Pearson correlation coefficient	0.7237995	4.700180e−01	204	5.061116e−02	−0.08668992	0.1860254

```
## To pdf-file also
# setwd("figures&tables/protocol-output")
# pdf("tumorsize-analysis-tables.pdf", width=15, height=9)
# grid.arrange(tableGrob(size_tests),
#              tableGrob(size_summary),
#              nrow=2, ncol=1)
# dev.off()
# for (i in 1:2) setwd("../")
```

6.6. Visualization

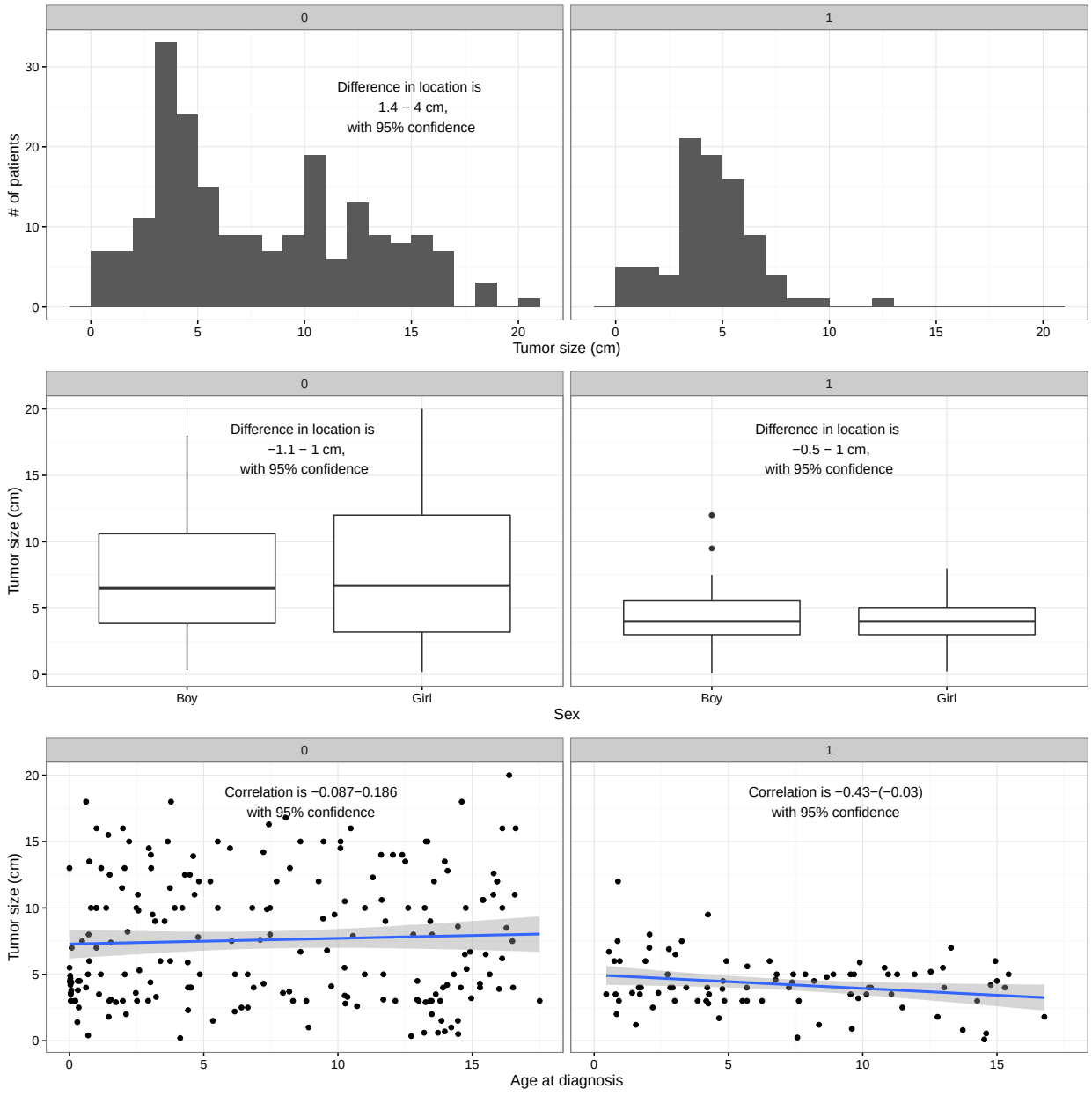
```
library(ggplot2)
options(scipen=999)
# setwd("figures&tables/protocol-output")

## 1. Plot: Tumor size by stage
ggplot(subset(size, !is.na(stage))) +
  geom_boxplot(aes(x=stage, y=tumorsize)) +
  geom_jitter(aes(x=stage, y=tumorsize),
              position=position_jitter(w=0.1),
              alpha=0.7) +
  annotate("text",
           label=paste("p = ",
                       round(size_tests[1, "p.value"], 6),
                       ", against 'all identical'",
                       sep=""),
           x=2.5,
           y=21) +
  theme_bw() +
  ylab("Tumor size (cm)") +
  xlab("Stage at diagnosis")
```




```
ggsave("tumorsize-stage-plot.pdf", width=8, height=10)

## 2. Plot: 1) Tumorsize by III/non-III, 2) sex, and 3) age at diagnosis
library(gridExtra)
## Subplot 1
plot1_annotation <- data.frame(label = paste("Difference in location is \n",
                                             round(size_tests[2, "conf.low"], 3),
                                             "-",
                                             round(size_tests[2, "conf.high"], 3),
                                             "cm,\nwith 95% confidence"),
                               tumorsize = 15,
                               iccc_cns = factor("0", levels=c("0", "1")),
                               y = 25)
plot1 <- ggplot(subset(size, !is.na(iccc_cns))) +
```

```
  geom_histogram(aes(x=tumorsize), binwidth=1) +
  facet_wrap(~ iccc_cns) +
  geom_text(data = plot1_annotation, aes(label = label, x=tumorsize, y=y)) +
  theme_bw() +
  ylab("# of patients") +
  xlab("Tumor size (cm)")
## Subplot 2
plot2_annotation <- data.frame(label = c(paste("Difference in location is \n",
                                               round(size_tests[4, "conf.low"], 3),
                                               "-",
                                               round(size_tests[4, "conf.high"], 3),
                                               "cm,\nwith 95% confidence"),
                                         paste("Difference in location is \n",
                                               round(size_tests[3, "conf.low"], 3),
                                               "-",
                                               round(size_tests[3, "conf.high"], 3),
                                               "cm,\nwith 95% confidence")),
                               x = c(1.5, 1.5),
                               y = c(17, 17),
                               iccc_cns = c("0", "1"))

plot2 <- ggplot(subset(size, !is.na(sex) & !is.na(iccc_cns))) +
  geom_boxplot(aes(x=sex, y=tumorsize)) +
  geom_text(data = plot2_annotation, aes(label = label, x=x, y=y)) +
  facet_wrap(~ iccc_cns) +
  theme_bw() +
  ylab("Tumor size (cm)") +
  xlab("Sex")
## Subplot 3
plot3_annotation <- data.frame(label = c(paste("Correlation is ",
                                               round(size_tests[5, "conf.low"], 3),
                                               "-",
                                               "(", round(size_tests[5, "conf.high"], 3),
                                               ")",
                                               "\nwith 95% confidence",
                                               sep=""),
                                         paste("Correlation is ",
                                               round(size_tests[6, "conf.low"], 3),
                                               "-",
                                               round(size_tests[6, "conf.high"], 3),
                                               "\nwith 95% confidence",
                                               sep="")),
                               x=c(9, 9),
                               y=c(18, 18),
                               iccc_cns=c("1", "0"))
plot3 <- ggplot(data = subset(size, !is.na(age_dx) & !is.na(iccc_cns)),
                aes(x=age_dx, y=tumorsize)) +
  geom_point() +
  facet_wrap(~ iccc_cns) +
  stat_smooth(method = "lm", formula = y ~ x) +
  theme_bw() +
  xlab("Age at diagnosis") +
  ylab("Tumor size (cm)") +
```

```
  geom_text(data = plot3_annotation,
            aes(label = label, x=x, y=y))

## Merge, save, and print
grid.arrange(plot1, plot2, plot3, nrow=3, ncol=1)
```




```
pdf("tumorsize-cns-sex-age-plot.pdf", width=9, height=13)
grid.arrange(plot1, plot2, plot3, nrow=3, ncol=1)
dev.off()
```

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

```
rm(plot1, plot2, plot3, plot1_annotation, plot2_annotation, plot3_annotation)
options(scipen=0)
# for(i in 1:2) setwd("../")
```

7. Batch 4: Transplantations

- What proportion of patients by ICCC3 -main groups was treated with autologous stem cell transplantation?

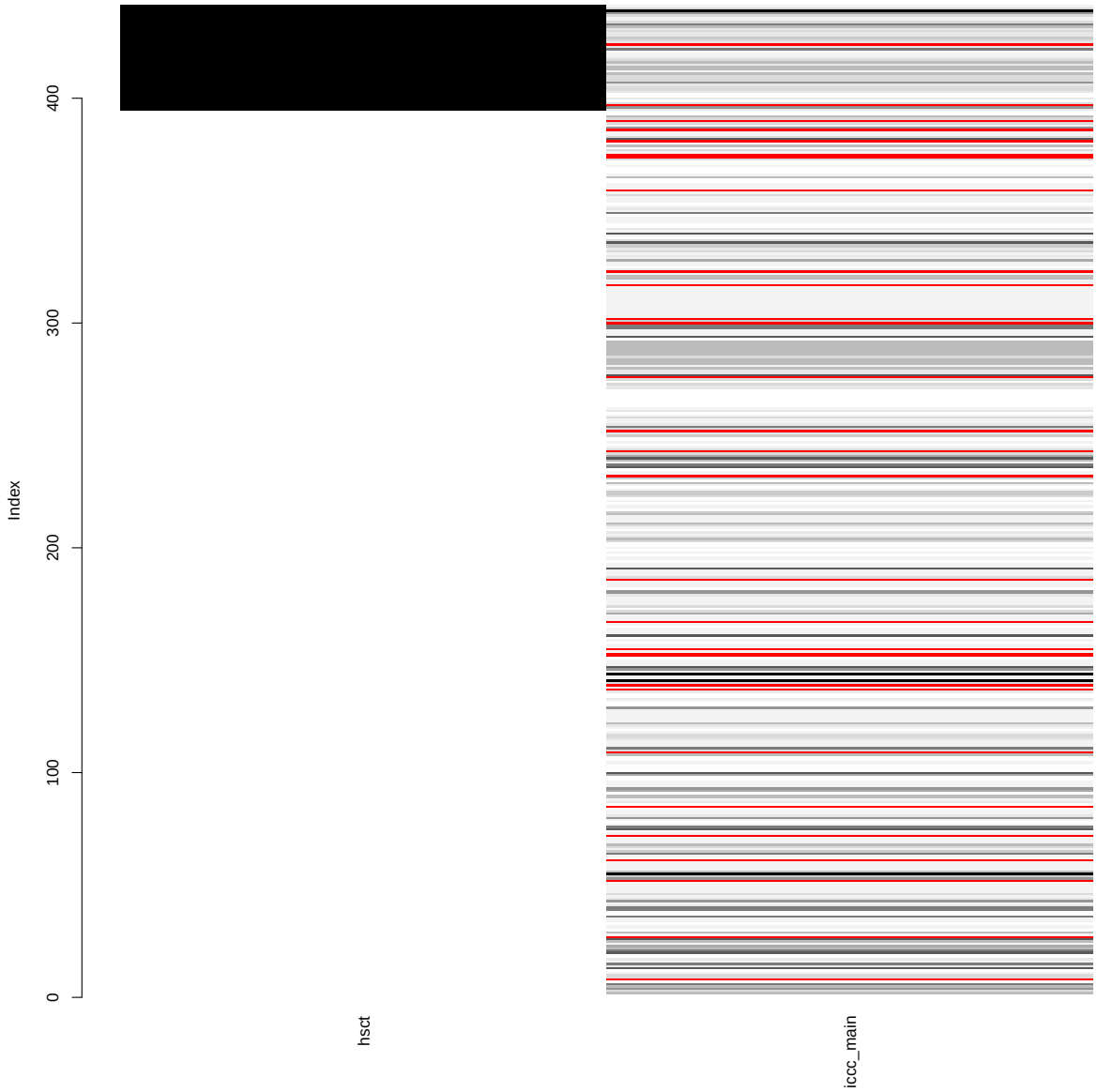
7.1. Selecting and filtering from clean data

```
library(dplyr)
hsct <- pottikk %>%
  select(hsct, hsct_place, iccc_main) %>%
  ## Filter transplants done in TAYS or unknown place in
  filter(hsct_place == "TAYS" | is.na(hsct_place)) %>%
  select(-hsct_place) %>%
  arrange(hsct) %>%
  mutate(iccc_main = factor(iccc_main))
```

7.2. Describing and Handling missing values

Complete cases: 417 out of 447 patients

```
library(VIM)
matrixplot(hsct)
```



Handling: complete-case analysis.

```
library(dplyr)
hsct <- hsct %>% filter(!is.na(iccc_main))
```

7.3. Analysis

Primary analytics methodology: Multiple exact binomial tests

```
library(broom)
library(dplyr)
hsct_tests <- data.frame()
frequencies <- tidy(table(hsct))
```

```r
for (i in unique(hsct$iccc_main)) {
  result <- tidy(binom.test(x = rev(frequencies$Freq[frequencies$iccc_main == i])))
  hsct_tests <- bind_rows(hsct_tests, bind_cols(data.frame(diagnosis = i), result))
}
rm(frequencies, result)
```

7.4. Tabulation

```r
library(gridExtra)
hsct_tests <- hsct_tests %>%
  rename(n_treated = statistic, n_all = parameter) %>%
  select(-p.value)
grid.arrange(tableGrob(hsct_tests))
```

	diagnosis	estimate	n_treated	n_all	conf.low	conf.high
1	II	0.05000000	4	80	0.013789394	0.12309874
2	VI	0.13513514	5	37	0.045371992	0.28774780
3	IX	0.28571429	12	42	0.157191467	0.44583880
4	VIII	0.23076923	3	13	0.050381073	0.53813154
5	X	0.12500000	2	16	0.015513604	0.38347624
6	III	0.03973510	6	151	0.014719454	0.08447802
7	XI	0.00000000	0	15	0.000000000	0.21801936
8	IV	0.30555556	11	36	0.163473985	0.48107063
9	VII	0.00000000	0	6	0.000000000	0.45925813
10	V	0.09090909	1	11	0.002298972	0.41277992
11	XII	0.25000000	1	4	0.006309463	0.80587955

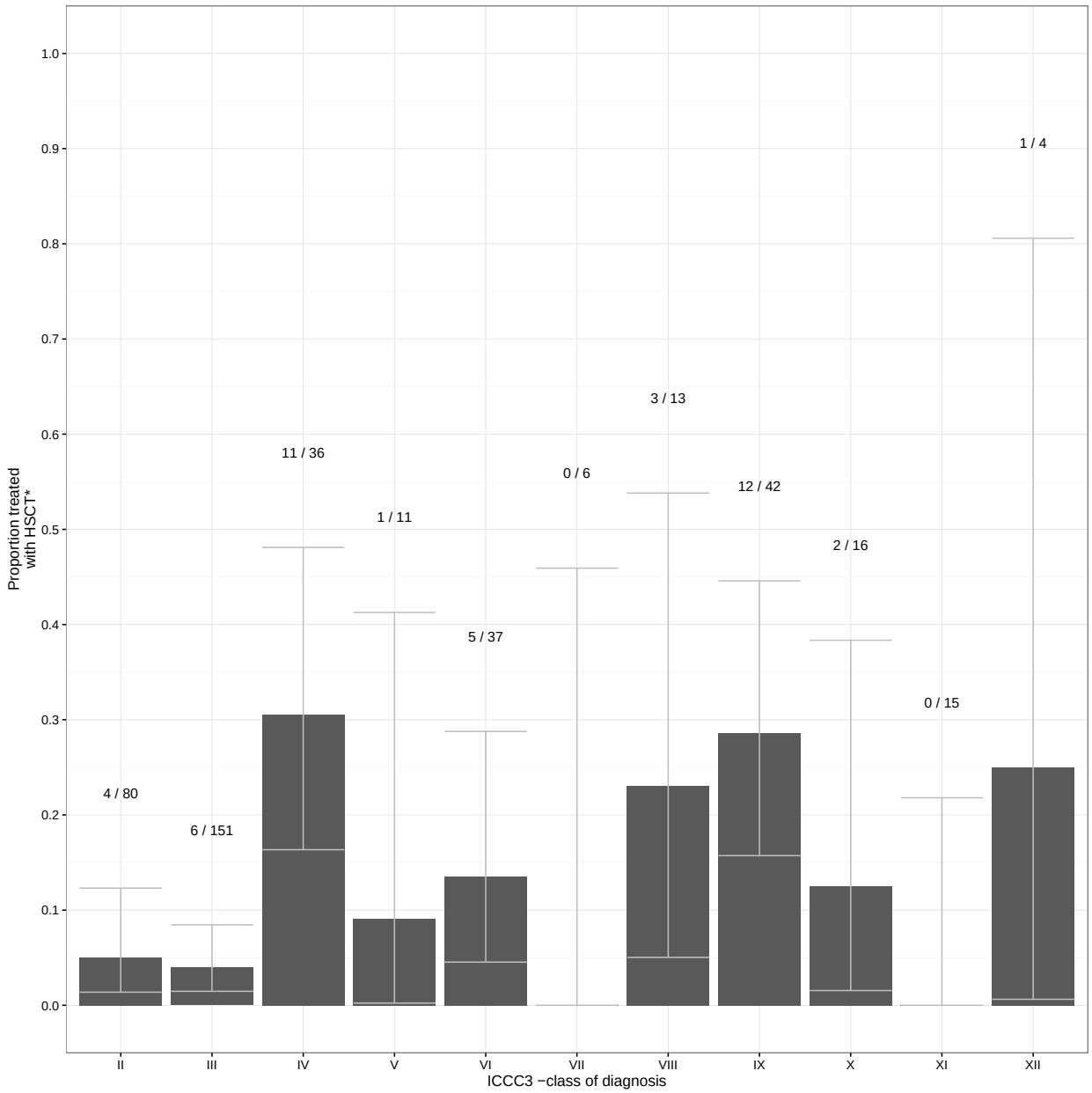
```
# setwd("figures&tables/protocol-output")
pdf("hsct-table.pdf")
grid.arrange(tableGrob(hsct_tests))
dev.off()
```

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

```
# for (i in 1:2) setwd("../")
```

7.5. Visualization

```
library(ggplot2)
## Names and ordering for the plot
if (is.numeric(hsct$hsct)) hsct$hsct <- factor(hsct$hsct)
hsct$iccc_main <- factor(hsct$iccc_main,
                         levels=c("II", "III", "IV", "V",
                                  "VI", "VII", "VIII",
                                  "IX", "X", "XI", "XII"))
names(hsct_tests) <- c("iccc_main", "estimate",
                       "n_treated", "n_all",
                       "conf.low", "conf.high")
hsct_tests$iccc_main <- factor(hsct_tests$iccc_main,
                         levels=c("II", "III", "IV", "V",
                                  "VI", "VII", "VIII",
                                  "IX", "X", "XI", "XII"))

## Bar graph with errorbars and  relative frequencies
hsct_plot <- ggplot(hsct_tests) +
  geom_bar(aes(x=iccc_main, y=estimate), stat="identity") +
  geom_errorbar(aes(x=iccc_main, ymin=conf.low, ymax=conf.high),
                color="grey") +
  geom_text(aes(label=paste(n_treated, "/", n_all),
                x=iccc_main,
                y=conf.high + 0.1)) +
  xlab("ICCC3 -class of diagnosis") +
  ylab("Proportion treated\nwith HSCT*") +
  scale_y_continuous(breaks=seq(0, 1, by=0.1), limits=c(0, 1)) +
  theme_bw()
hsct_plot
```




```
# setwd("figures&tables/protocol-output")
ggsave("transplantations-plot.pdf", width=8, height=10)
# for (i in 1:2) setwd("../")
```

8. Batch 5: Diagnostic imaging trends

- What were the trends of diagnostic imaging modalities like?
- How likely was there a change in the populations of trends of diagnostic imaging modalities?

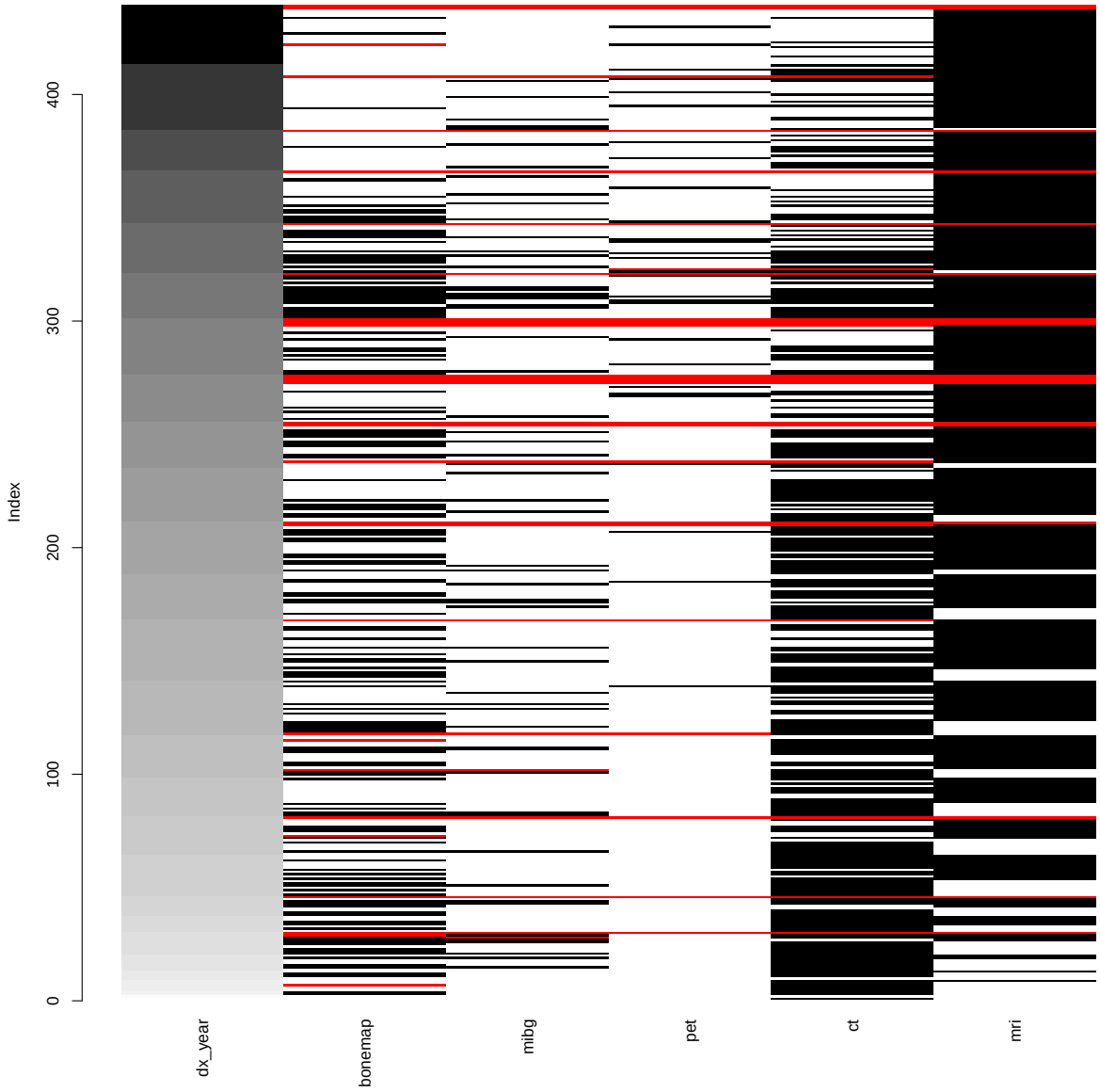
8.1. Selecting and filtering from clean data

```
library(dplyr)
imaging <- pottikk %>%
  select(dx_year, bonemap:mri) %>%
  ## Delete 2015 (not full year of data)
  filter(dx_year < 2015) %>%
  arrange(dx_year, mri)
```

8.2. Describing and Handling missing values

Missingness structure:

```
library(VIM)
matrixplot(imaging)
```




```
## Delete cases with all imaging data missing
imaging <- imaging %>%
  filter(any(!is.na(bonemap),
             !is.na(mibg),
             !is.na(pet),
             !is.na(ct),
             !is.na(mri)))
## Calculate missing values in variables
imaging_missing <- apply(imaging, 2, function(x) sum(is.na(x)))
imaging_missing
```

```
## dx_year bonemap    mibg     pet      ct     mri
##       0      31      27      26      25      19
```

Handling: Complete case analyses by imaging variable.

8.3. Analysis

8.3.1. Description/Summarization

```
imaging_summary <- imaging %>%
  group_by(dx_year) %>%
  summarise_each(funs(mean(., na.rm=TRUE), sum(., na.rm=TRUE)))
```

- Tabulation:

```
library(gridExtra)
grid.arrange(tableGrob(round(imaging_summary, 4)))
```

	dx_year	bonemap_mean	mibg_mean	pet_mean	ct_mean	mri_mean	bonemap_sum	mibg_sum	pet_sum	ct_sum	mri_sum
1	1987	0	0	0	1	0	0	0	0	1	0
2	1988	0	0	0	0	0	0	0	0	0	0
3	1990	1	0	0	1	0	2	0	0	2	0
4	1991	0	0	0	1	0.2	0	0	0	5	1
5	1992	0.5	0	0	0.75	0.25	2	0	0	3	1
6	1993	0.4286	0.2857	0	1	0.2857	3	2	0	7	2
7	1994	0.875	0.5	0	0.8889	0.3333	7	4	0	8	3
8	1995	0.4286	0	0	1	0.5714	3	0	0	7	4
9	1996	0.625	0.25	0	0.75	0.5	5	2	0	6	4
10	1997	0.4444	0.0556	0	0.8889	0.6111	8	1	0	16	11
11	1998	0.4	0.0625	0	0.6875	0.5625	6	1	0	11	9
12	1999	0.2941	0.1176	0	0.7647	0.6471	5	2	0	13	11
13	2000	0.4118	0.1667	0	0.6842	0.7895	7	3	0	13	15
14	2001	0.4348	0.1739	0.0435	0.7083	0.75	10	4	1	17	18
15	2002	0.4231	0.0769	0	0.6154	0.8148	11	2	0	16	22
16	2003	0.35	0.2	0.05	0.75	0.75	7	4	1	15	15
17	2004	0.4762	0.0952	0.0476	0.8571	0.9091	10	2	1	18	20
18	2005	0.2917	0.125	0	0.7083	0.875	7	3	0	17	21
19	2006	0.5294	0.2353	0.0588	0.7647	0.8889	9	4	1	13	16
20	2007	0.2353	0.0588	0.1765	0.3529	1	4	1	3	6	17
21	2008	0.381	0.0952	0.0952	0.4286	1	8	2	2	9	22
22	2009	0.8421	0.3684	0.2105	0.7895	1	16	7	4	15	19
23	2010	0.5714	0.1905	0.25	0.65	0.9524	12	4	5	13	20
24	2011	0.4091	0.1818	0.0909	0.3182	1	9	4	2	7	22
25	2012	0.0588	0.1176	0.1176	0.5294	1	1	2	2	9	17
26	2013	0.0357	0.1786	0.1429	0.4286	0.9655	1	5	4	12	28
27	2014	0.087	0	0.0833	0.1667	1	2	0	2	4	24

```
# setwd("figures&tables/protocol-output")
pdf("imaging-summary-table.pdf", width=14, height=12)
grid.arrange(tableGrob(round(imaging_summary, 4)))
dev.off()
```

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

```
# for (i in 1:2) setwd("../")
```

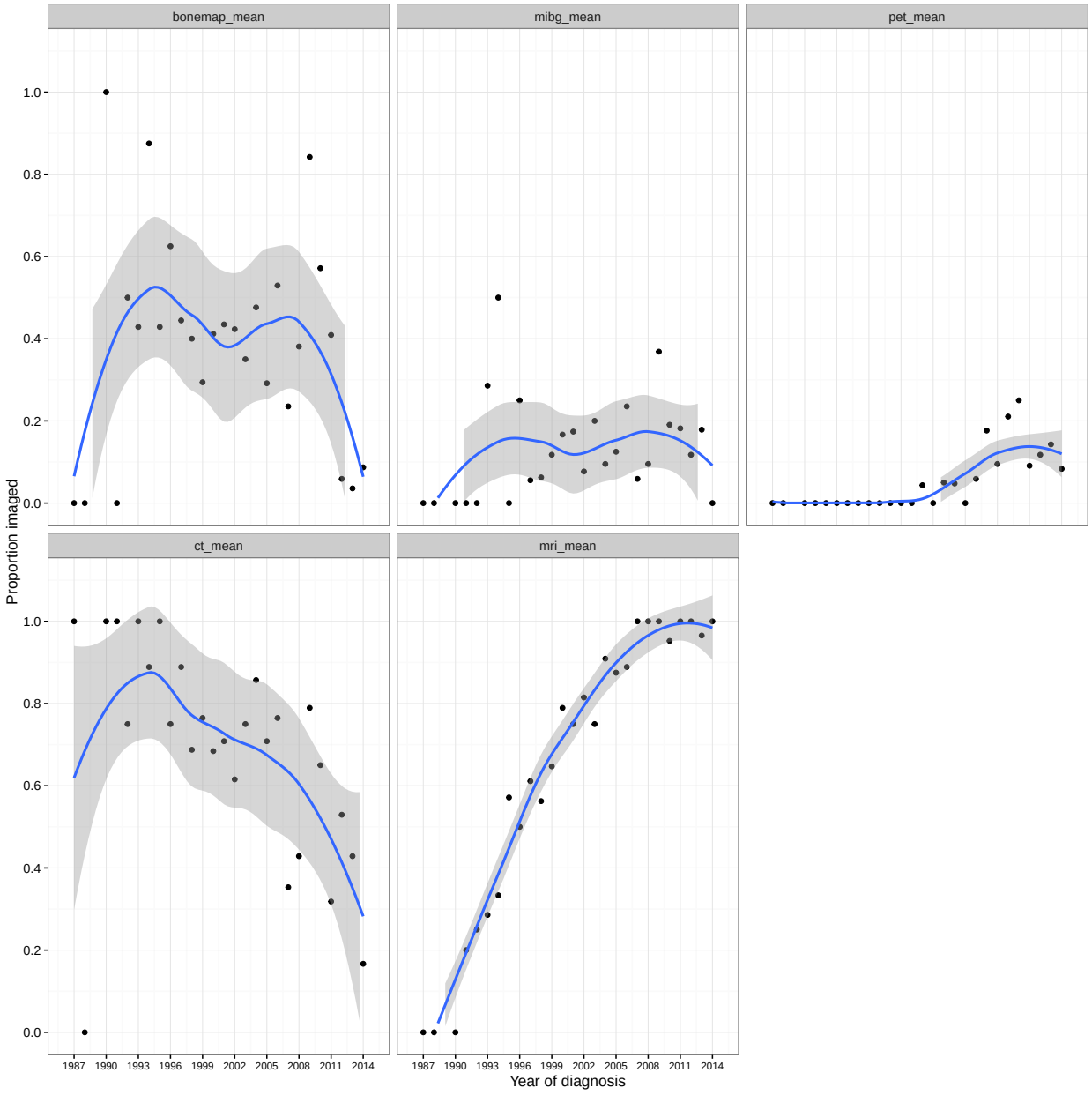
8.3.2. Analysis

Primary time-series analytics method:

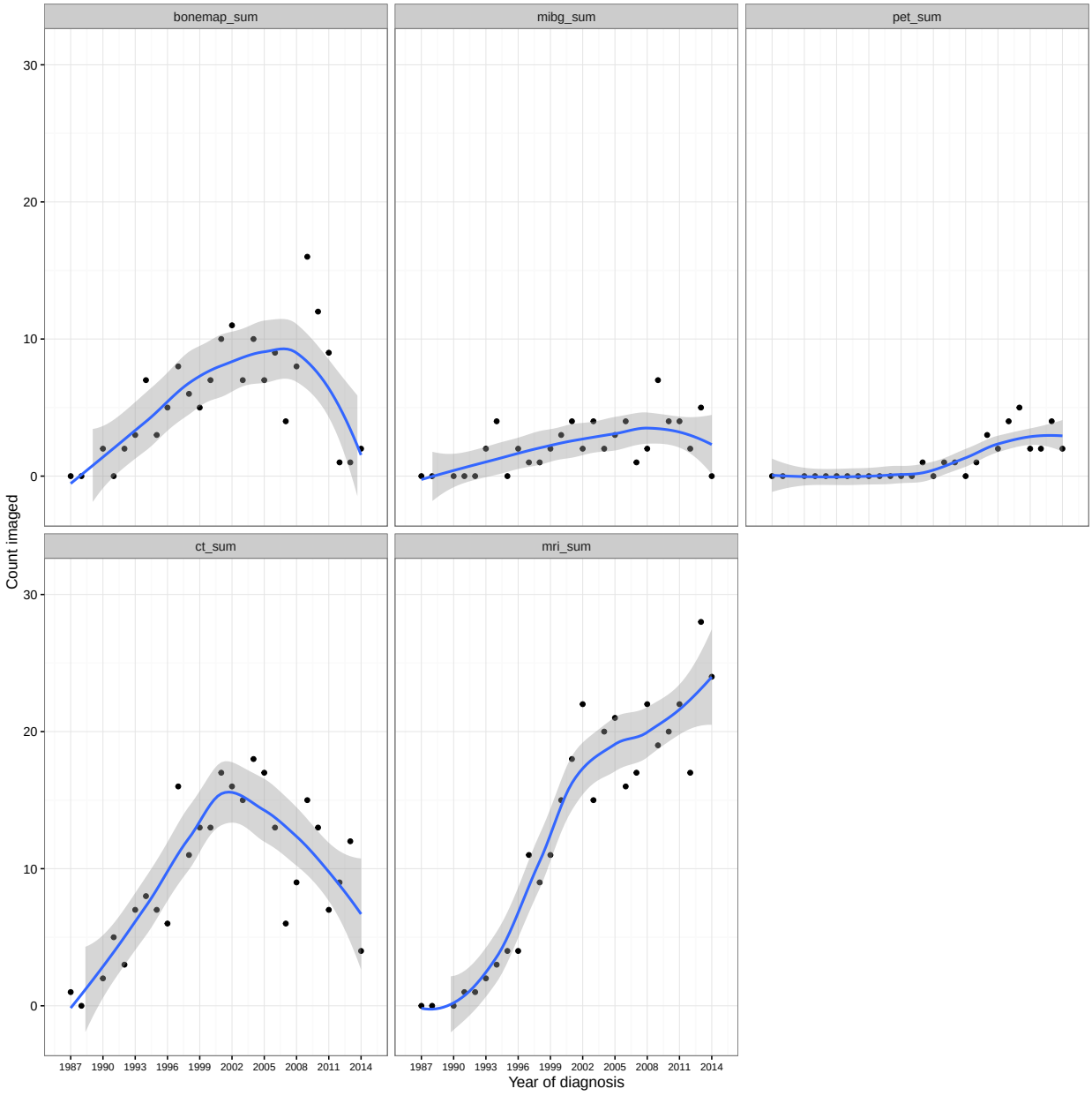
- Visualization and loess-smoother

Visualization:

```
library(ggplot2)
## Goal: Modality-faceted time-series scatter with smoothers
## 1. Wrangle data for handier plotting
library(tidyr)
library(dplyr)
imaging_prop <- imaging_summary %>%
  select(dx_year, contains("mean")) %>%
  gather(modality, prop, contains("mean"))
imaging_count <- imaging_summary %>%
  select(dx_year, contains("sum")) %>%
  gather(modality, count, contains("sum"))
## Proportion time-series plot
imaging_prop_plot <- ggplot(imaging_prop) +
  geom_point(aes(x=dx_year, y=prop)) +
  geom_smooth(aes(x=dx_year, y=prop), method="loess") +
  facet_wrap(~ modality) +
  scale_y_continuous(breaks=seq(0, 1, by=0.2), limits=c(0, 1.1)) +
  scale_x_continuous(breaks=seq(1987, 2014, by=3), limits=c(1986, 2015)) +
  theme_bw() +
  ylab("Proportion imaged") +
  xlab("Year of diagnosis") +
  theme(axis.text.x = element_text(size=8))
## Count time-series plot
imaging_count_plot <- ggplot(imaging_count) +
  geom_point(aes(x=dx_year, y=count)) +
  geom_smooth(aes(x=dx_year, y=count), method="loess") +
  facet_wrap(~ modality) +
  scale_y_continuous(breaks=seq(0,30, by=10), limits=c(-2, 31)) +
  scale_x_continuous(breaks=seq(1987, 2014, by=3), limits=c(1986, 2015)) +
  theme_bw() +
  xlab("Year of diagnosis") +
  ylab("Count imaged") +
  theme(axis.text.x = element_text(size=8))
## Print and save
# setwd("figures&tables/protocol-output")
imaging_prop_plot
```




```
ggsave("imaging-trend-prop.pdf", width=9, height=8)
imaging_count_plot
```




```
ggsave("imaging-trend-count.pdf", width=9, height=8)
# for (i in 1:2) setwd("../")
rm(imaging_prop, imaging_count)
```

9. Batch 6: Metastasis proportions

- What was the proportion of metastasised cancers (at diagnosis) in...

a. III group?
b. non-III group?

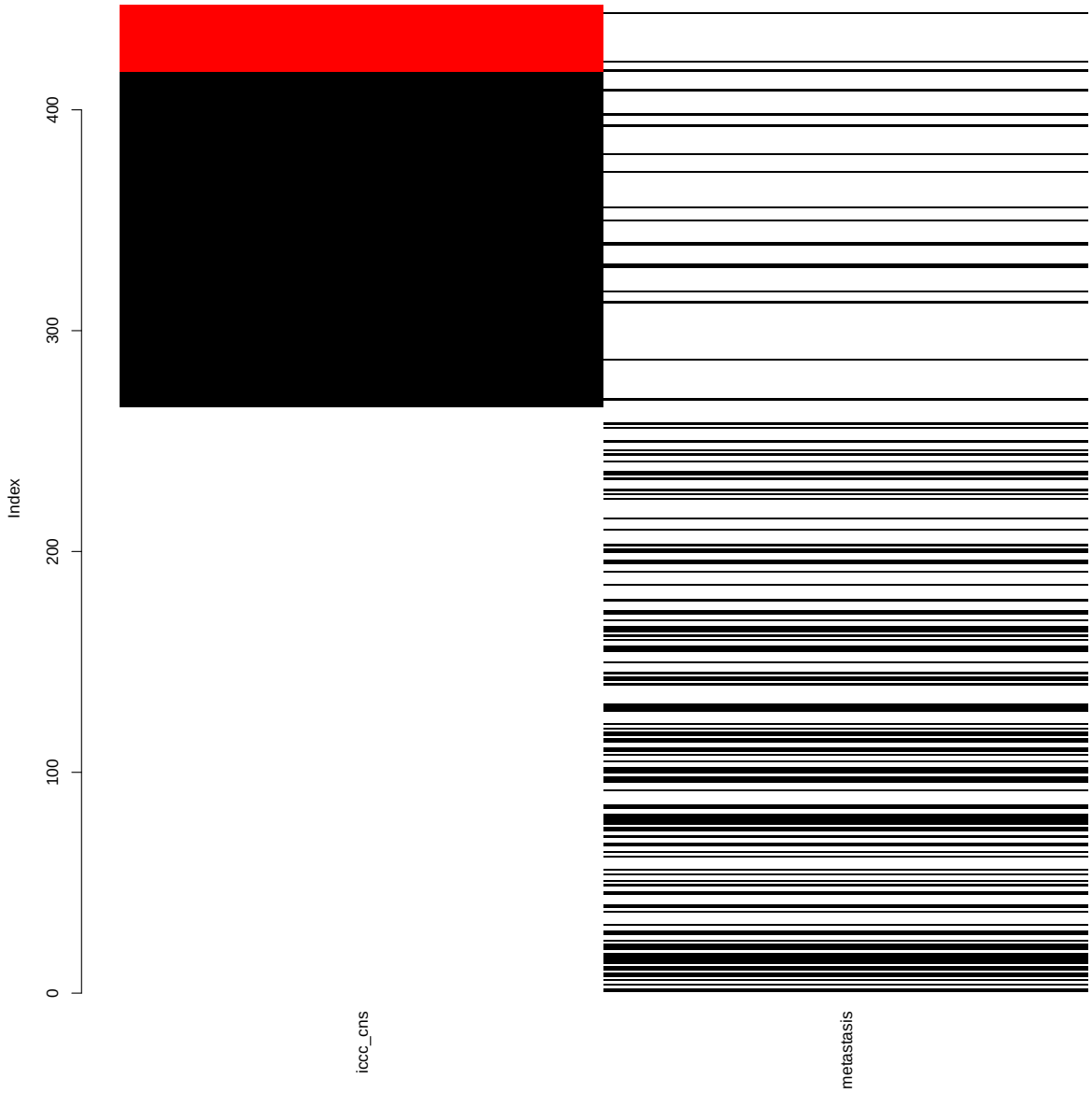
9.1. Calculating and merging metastasis-variable from raw data

```
library(dplyr)
library(readxl)
raw <- read_excel(path = path_raw)
metastasis <- raw %>%
  select(id = Id, ds = Ds,
         brain = MetastaasiAivot_KKdiagnoosivaihe,
         lung = MetastaasiKeuhkot_KKdiagnoosivaihe,
         bone = MetastaasiLuusto_KKdiagnoosivaihe,
         marrow = MetastaasiLuuydin_KKdiagnoosivaihe,
         liver = MetastaasiMaksa_KKdiagnoosivaihe,
         other = MetastaasiMuu_KKdiagnoosivaihe) %>%
  filter(ds == 1) %>%
  mutate(id = as.character(id))
## Metastasis variable:
## if any location-specific variable "Yes" (Kyllä) or
## locations are given in free text variable; gets value 1,
## otherwise no metastasis and gets 0.
metastasis$metastasis <- c()
for (i in 1:nrow(metastasis)) {
  metastasis_test <- with(metastasis[i, ],
                          brain == "Kyllä" & !is.na(brain) |
                            lung == "Kyllä" & !is.na(lung) |
                            bone == "Kyllä" & !is.na(bone) |
                            marrow == "Kyllä" & !is.na(marrow) |
                            liver == "Kyllä" & !is.na(liver) |
                            !is.na(other))
  if (metastasis_test) {
    metastasis$metastasis[i] <- 1
  } else {
    metastasis$metastasis[i] <- 0
  }
}
metastasis_i <- metastasis %>% select(id, metastasis)
## Merge metastasis data with ICCC-diagnosis class data
metastasis <- left_join(pottikk[ ,c("id", "iccc_cns")],
                        metastasis_i,
                        by="id")
metastasis <- metastasis %>%
  select(-id) %>%
  arrange(iccc_cns)
```

9.2. Describing and Handling missing values

Missingness structure:

```
library(VIM)
matrixplot(metastasis)
```



- Missing values in CNS-tumors: 3
- Missing values in non-CNS-tumors: 27

Handling: Complete observations only.

```
metastasis <- metastasis[complete.cases(metastasis), ]
```

9.3. Analysis

9.3.1. Description/Summarising

```
metastasis_summary <- metastasis %>%
  group_by(iccc_cns) %>%
  summarise(all = n(), count = sum(metastasis),
            proportion = round(mean(metastasis), 6))
```

9.3.2. Inference

Primary analytics methodology: Two-sample test for equality of proportions with continuity correction.

```
library(broom)
metastasis_inference <- tidy(prop.test(x = metastasis_summary$count,
                                       n = metastasis_summary$all))
metastasis_inference <- cbind(
  data.frame(method = "2-sample approx. test\nof proportions\nwith continuity correction"),
  round(metastasis_inference, 6))
```

9.4. Tabulation

```
library(gridExtra)
# setwd("figures&tables/protocol-output")
grid.arrange(tableGrob(metastasis_summary),
             tableGrob(metastasis_inference))
```

	iccc_cns	all	count	proportion
1	0	265	102	0.384906
2	1	152	15	0.098684

	method	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
1	2−sample approx. test of proportions with continuity correction	0.384906	0.098684	37.79816	0	1	0.20568	0.366763

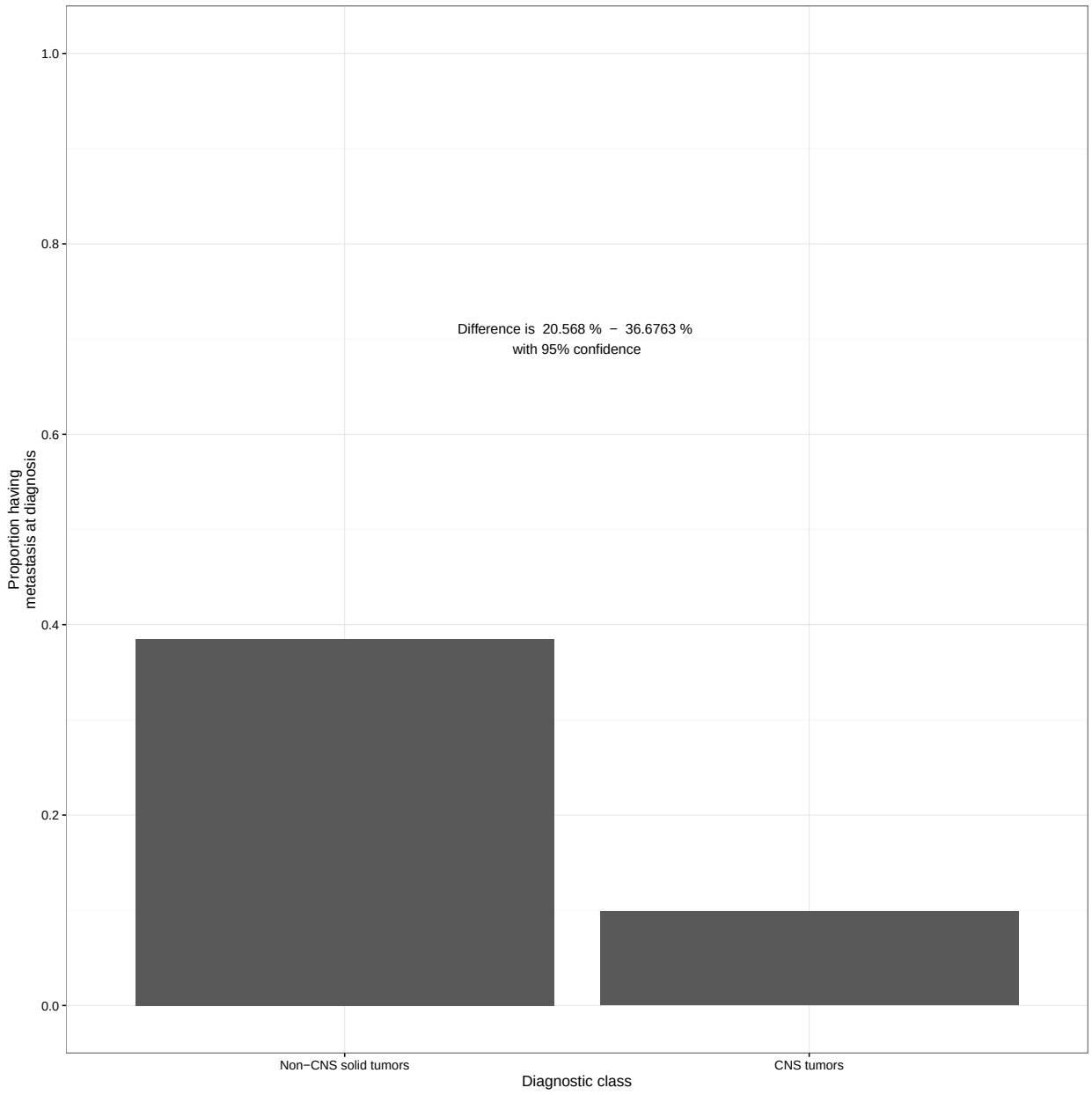
```
pdf("metastasis-table.pdf", width=10, height=8)
grid.arrange(tableGrob(metastasis_summary),
             tableGrob(metastasis_inference))
dev.off()
```

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

```
# for (i in 1:2) setwd("../")
```

9.5. Visualization

```
library(ggplot2)
metastasis_plot <- ggplot(metastasis_summary) +
  geom_bar(aes(x=as.character(iccc_cns), y=proportion), stat="identity") +
  scale_y_continuous(limits=c(0, 1), breaks=seq(0, 1, by=0.2)) +
  theme_bw() +
  annotate("text",
           label=paste("Difference is ",
                       metastasis_inference[1, "conf.low"]*100,
                       "%",
                       " - ",
                       metastasis_inference[1, "conf.high"]*100,
                       "%",
                       "\nwith 95% confidence"),
           x=1.5,
           y=0.7) +
  ylab("Proportion having\nmetastasis at diagnosis") +
  xlab("Diagnostic class") +
  scale_x_discrete(labels=c("Non-CNS solid tumors",
                            "CNS tumors"))
# setwd("figures&tables/protocol-output")
metastasis_plot
```




```
ggsave("metastasis-plot.pdf")
# for (i in 1:2) setwd("../")
```

10. Batch 7: Exploratory Cox regression modeling of survival time

10.1. Additional variables from the raw data

- Height, weight, and metastasis variables are added to the dataset

```
library(dplyr)
library(readxl)
```

```
cox_dataset <- pottikk

path_raw <- 'G:/Projects/potti/data/raw/potti-kk-20150529.xlsx'
raw <- read_excel(path = path_raw)
weightheight <- raw %>%
  select(id = Id, ds = Ds,
         weight = Paino_KKyleinen,
         height = Pituus_KKyleinen) %>%
  mutate(id = as.character(id)) %>%
  filter(ds == 1) %>%
  select(-ds)

# Join datasets and select potentially useful explanatory variables
# Filter to non-CNS tumours
# Categorise 4-class stage to 2-class: stage 4 vs. other classes

cox_dataset <- cox_dataset %>%
  left_join(weightheight, by="id") %>%
  left_join(metastasis_i, by="id") %>%
  filter(iccc_cns == 0) %>%
  select(c(-(1:3), -5, -7, -9, -(16:21), -24)) %>%
  mutate(stage = ifelse(stage == "4", 1, 0),
         sex = ifelse(sex == "Girl", 1, 0))

# Fix entry errors in weight and height variables
cox_dataset$weight[which(cox_dataset$weight == 2780)] <- 2.780
cox_dataset$height[which(cox_dataset$height == 17.3)] <- 173

# Body area as a composite measure of body size by
# Du Bois -formula: = 0.007184 x height(m)^0.725 x weight(kg)^0.425, and
# Delete height and weight from the dataset
cox_dataset <- cox_dataset %>%
  mutate(area = 0.007184 * height^0.725 * weight^0.425) %>%
  select(-weight, -height)
```

10.2. Cox proportional hazards regression with all predictor combinations

10.2.1. Modeling

- Severe data dregding if not interpreted and reported appropriately

```
# Generate all possible formulas (= response-predictors -variations)
all_predictors <- colnames(cox_dataset)[-(2:3)]

library(gtools)
formulas <- c()
for (r in 1:length(all_predictors)) {
combs <- combinations(n = length(all_predictors), r = r, v = all_predictors)
forms <- apply(combs, 1, function(x) {
  preds <- paste(x, collapse = " + ")
  paste("Surv(followup, death)" , preds, sep = " ~ ")})
formulas <- c(formulas, forms)
```

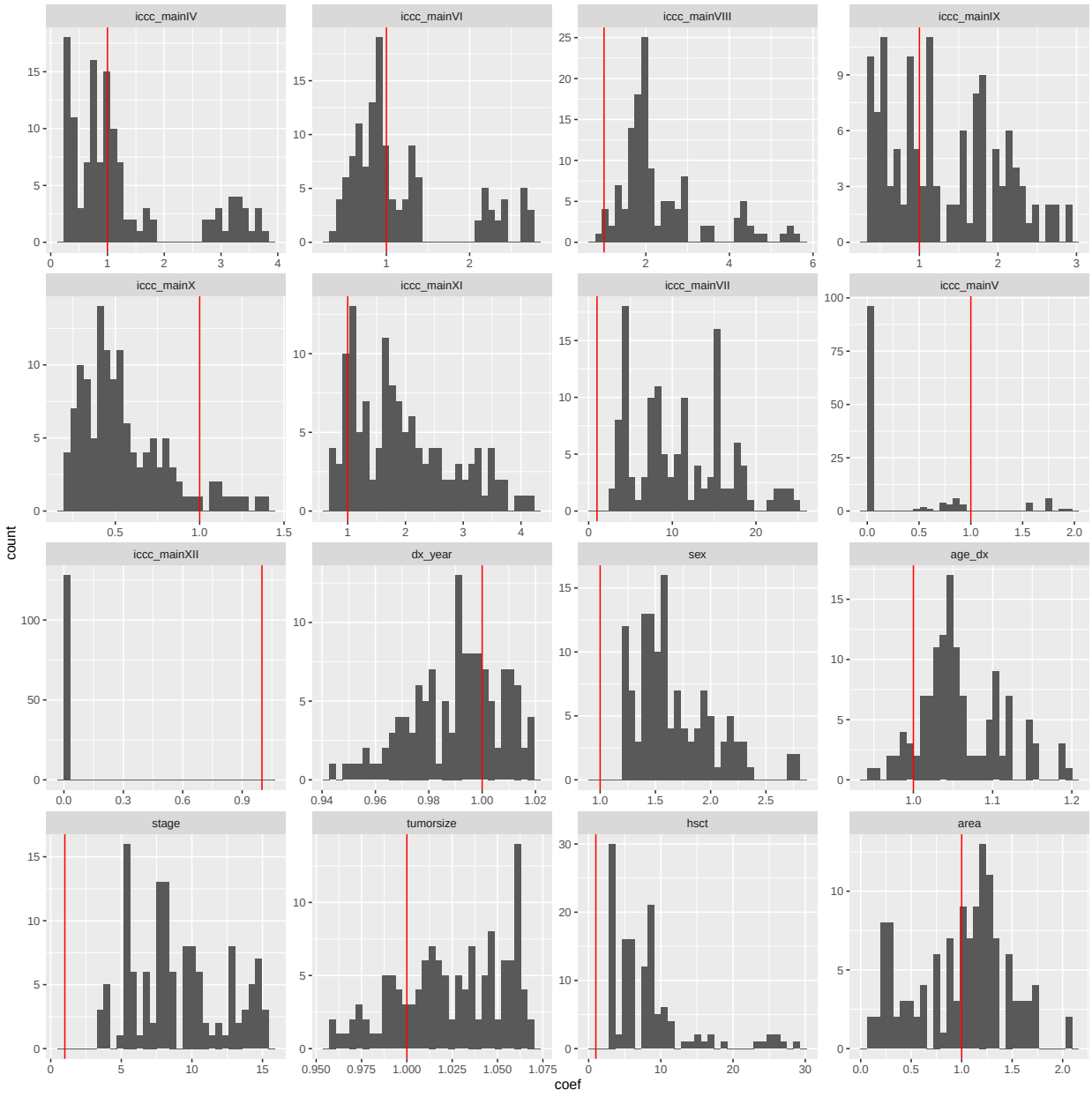
```
}

# Run Cox regression with each formula,
# Do diagnostics for the model, and
# Collect the relevant statistics to results data.frame
library(survival)
library(broom)

# 2 formula-columns, 21 coefficients, and 8 model statistics
cox_res <- matrix(nrow=length(formulas), ncol=27)

for (i in 1:length(formulas)) {

  # Model fit
  fit <- coxph(as.formula(formulas[i]), data = cox_dataset)

  # Set column names in the first run
  if (i == 1) {
    ests <- tidy(fit)
    mod_stats <- glance(fit)
    colnames(cox_res) <- c("model",
                           "n_predictors",
                           paste("iccc_main", unique(cox_dataset$iccc_main), sep=""),
                           all_predictors[-1],
                           colnames(mod_stats)[c(4, 6, 8, 9, 10, 14, 15)],
                           "ph_assum_pvalue")
  }

  # Add formula
  cox_res[i, 1] <- formulas[i]
  # How many predictors
  library(stringr)
  cox_res[i, 2] <- str_count(formulas[i], "\\+") + 1

  # Add predictor statistics
  ests <- tidy(fit)
  for (b in 1:nrow(ests)) {
    cox_res[i, which(colnames(cox_res) == ests$term[b])] <- as.character(round(ests$estimate[b], 5))
  }

  # Model statistics
  mod_stats <- glance(fit)[c(4, 6, 8, 9, 10, 14, 15)]
  mod_stats <- apply(mod_stats, 2, function(x) {as.character(x)})
  cox_res[i, 20:26] <- mod_stats

  # Test for proportional hazards (model-level)
  zph_test <- cox.zph(fit)$table
  zph_test_pvalue <-  zph_test[nrow(zph_test), ncol(zph_test)]
  cox_res[i, 27] <- as.character(zph_test_pvalue)
}
# iccc_mainII is the reference class for other classes
cox_res[ , 3] <- rep("ref", 255)
```

```r
cox_res <- data.frame(cox_res)
cox_res[ , c(-1, -3)] <- apply(cox_res[ , c(-1, -3)], 2, as.numeric)
cox_res$model <- as.character(cox_res$model)

# Coefficient transformation: log(HR) -> HR
cox_res[ , 4:19] <- apply(cox_res[ , 4:19], 2, exp)
```

10.2.2. Visualization of the results

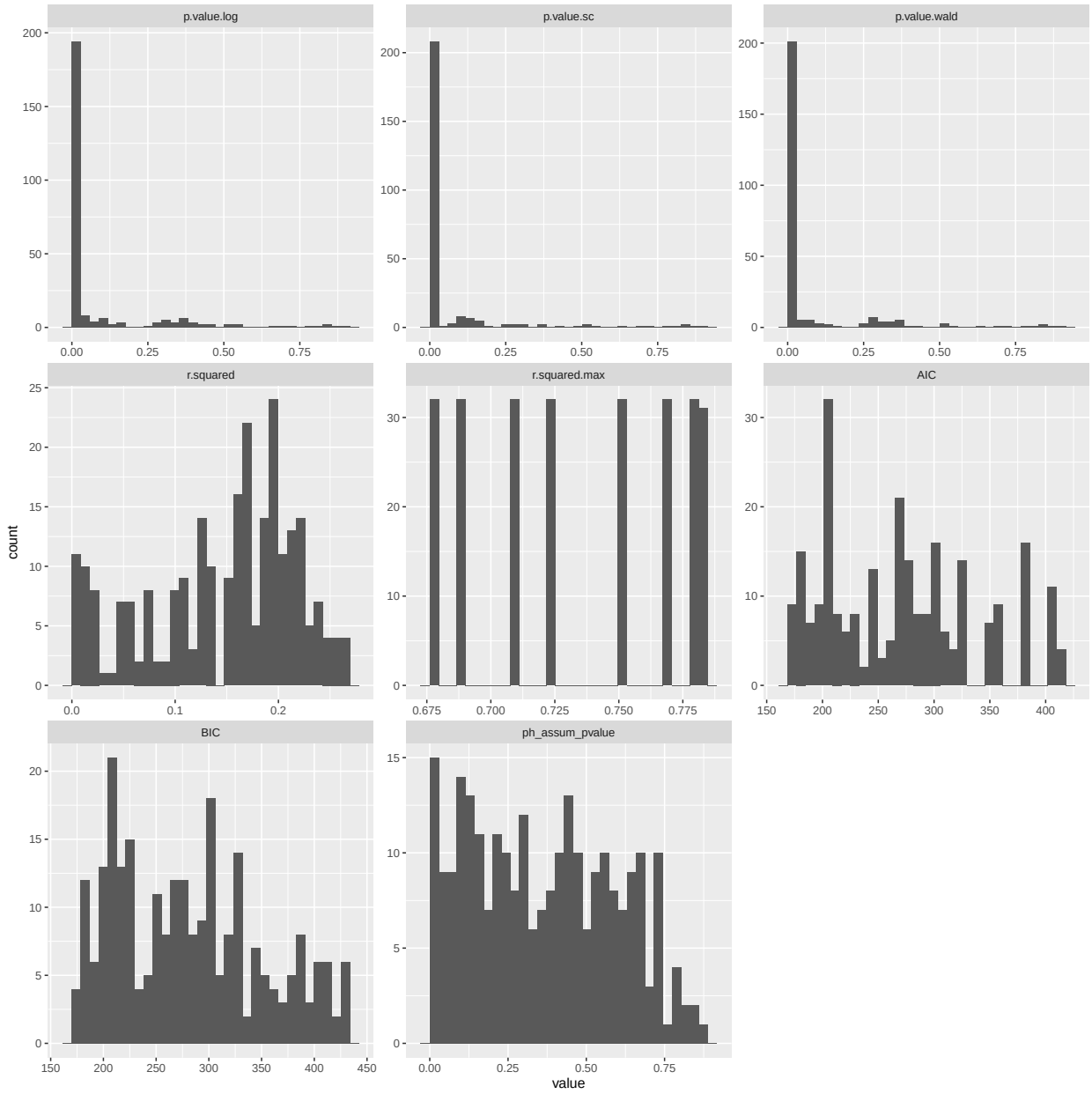
Coefficients:

```r
library(ggplot2)
library(dplyr)
library(tidyr)

cox_res_vis_coef <- cox_res %>%
  gather("predictor", "coef", 4:19)

ggplot(cox_res_vis_coef) +
  geom_histogram(aes(x = coef)) +
  facet_wrap(~ predictor, scales="free") +
  geom_vline(xintercept = 1, color="red")
```



Goodness-of-fit -statistics:

```
library(ggplot2)
library(dplyr)
library(tidyr)

cox_res_vis_stats <- cox_res %>%
  select(20:27) %>%
  gather("stat", "value")

ggplot(cox_res_vis_stats) +
  geom_histogram(aes(x=value)) +
  facet_wrap(~ stat, scales="free")
```



"The best fit" models (20):

```
library(gridExtra)
library(dplyr)

cox_res %>%
  filter(ph_assum_pvalue > 0.05 &
           p.value.log < 0.05 &
           p.value.sc < 0.05 &
           p.value.wald < 0.05) %>%
  arrange(desc(r.squared), desc(AIC), desc(BIC)) %>%
  slice(1:20) %>%
  select(model)
```

model

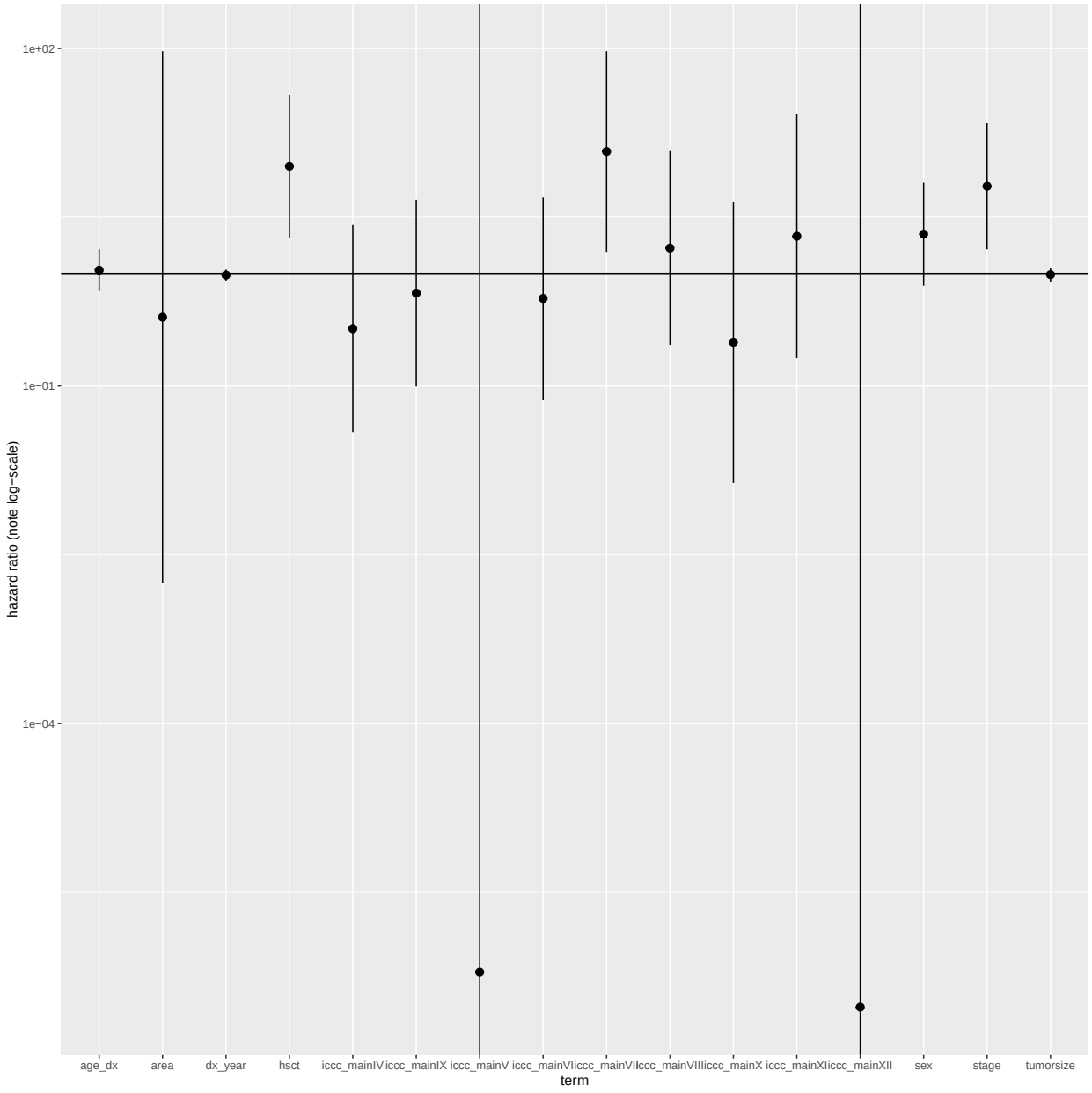
Surv(followup, death) ~ age_dx + area + dx_year + hsct + iccc_main + sex + stage
Surv(followup, death) ~ age_dx + area + hsct + iccc_main + sex + stage
Surv(followup, death) ~ area + dx_year + hsct + iccc_main + sex + stage
Surv(followup, death) ~ area + hsct + iccc_main + sex + stage
Surv(followup, death) ~ age_dx + area + dx_year + hsct + iccc_main + stage
Surv(followup, death) ~ age_dx + area + hsct + iccc_main + stage
Surv(followup, death) ~ area + dx_year + hsct + iccc_main + stage
Surv(followup, death) ~ area + hsct + iccc_main + stage
Surv(followup, death) ~ age_dx + area + dx_year + hsct + iccc_main + sex + stage + tumorsize
Surv(followup, death) ~ area + dx_year + hsct + iccc_main + sex + stage + tumorsize
Surv(followup, death) ~ age_dx + area + hsct + iccc_main + sex + stage + tumorsize
Surv(followup, death) ~ area + hsct + iccc_main + sex + stage + tumorsize
Surv(followup, death) ~ age_dx + area + dx_year + hsct + iccc_main + stage + tumorsize
Surv(followup, death) ~ age_dx + area + hsct + iccc_main + stage + tumorsize
Surv(followup, death) ~ age_dx + dx_year + hsct + iccc_main + sex + stage
Surv(followup, death) ~ area + dx_year + hsct + iccc_main + stage + tumorsize
Surv(followup, death) ~ age_dx + hsct + iccc_main + sex + stage
Surv(followup, death) ~ dx_year + hsct + iccc_main + sex + stage
Surv(followup, death) ~ hsct + iccc_main + sex + stage
Surv(followup, death) ~ area + hsct + iccc_main + stage + tumorsize

10.3. Exploring individual models and associations

10.3.1. Full model

The full model (all initially selected predictors):

```
library(broom)
library(survival)
library(dplyr)
library(ggplot2)

cox_dataset %>%
  coxph(formula = Surv(followup, death) ~ age_dx + area + dx_year +
                  hsct + iccc_main + sex + stage + tumorsize,
        data = .) %>%
  tidy() %>%
  ggplot() +
    geom_pointrange(aes(x = term,
                        y = exp(estimate),
                        ymin = exp(conf.low),
                        ymax = exp(conf.high))) +
    scale_y_log10() +
    geom_hline(yintercept=1) +
    labs(y = "hazard ratio (note log-scale)")
```



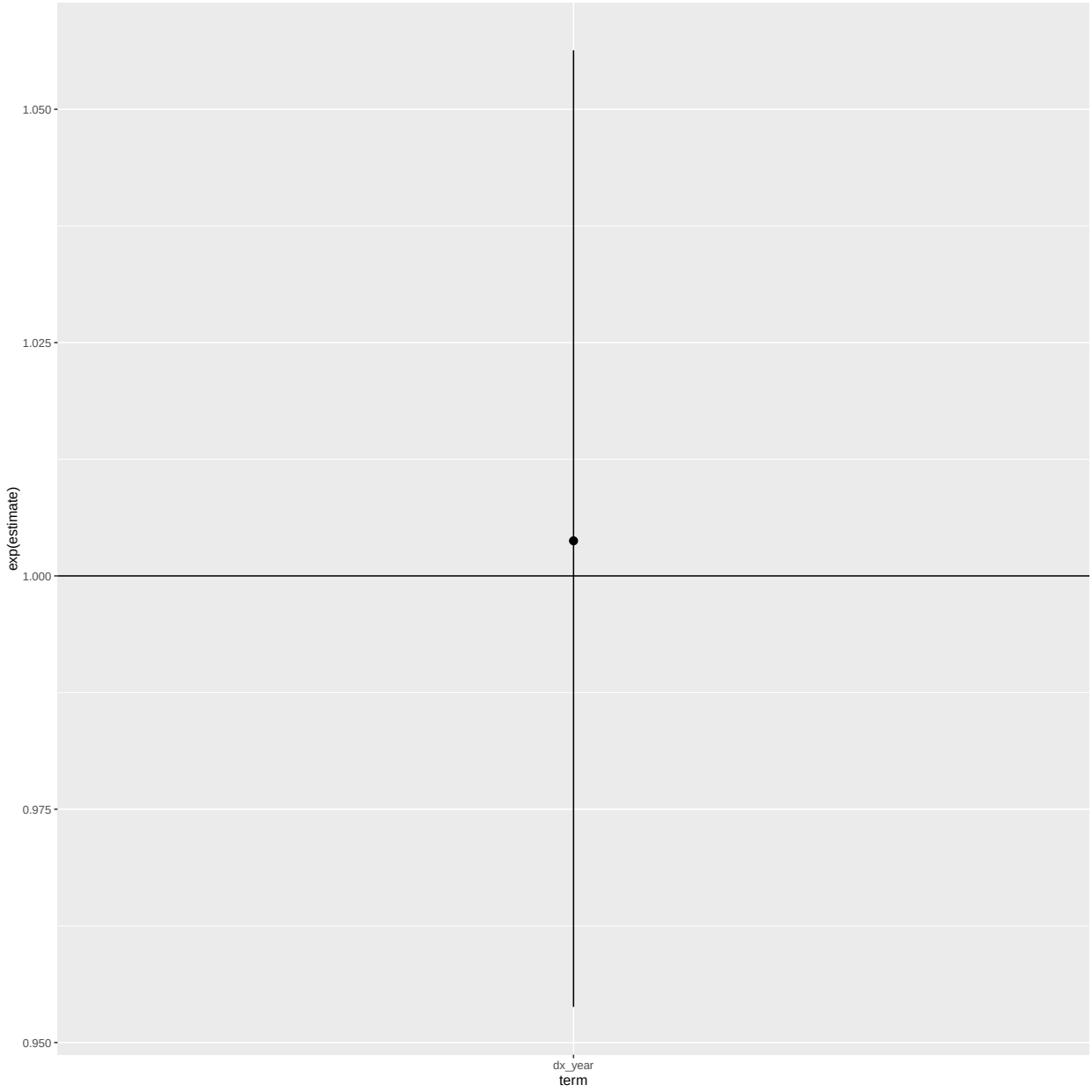
- ICCC-variable gives extremely imprecise estimates due to low sample size
- Given other variables, age, diagnosis year, and tumor size seem to be quite precisely close to no association (HR = 1)

10.3.2. Adjusted association between diagnosis year and survival?

When we model only diagnosis year:

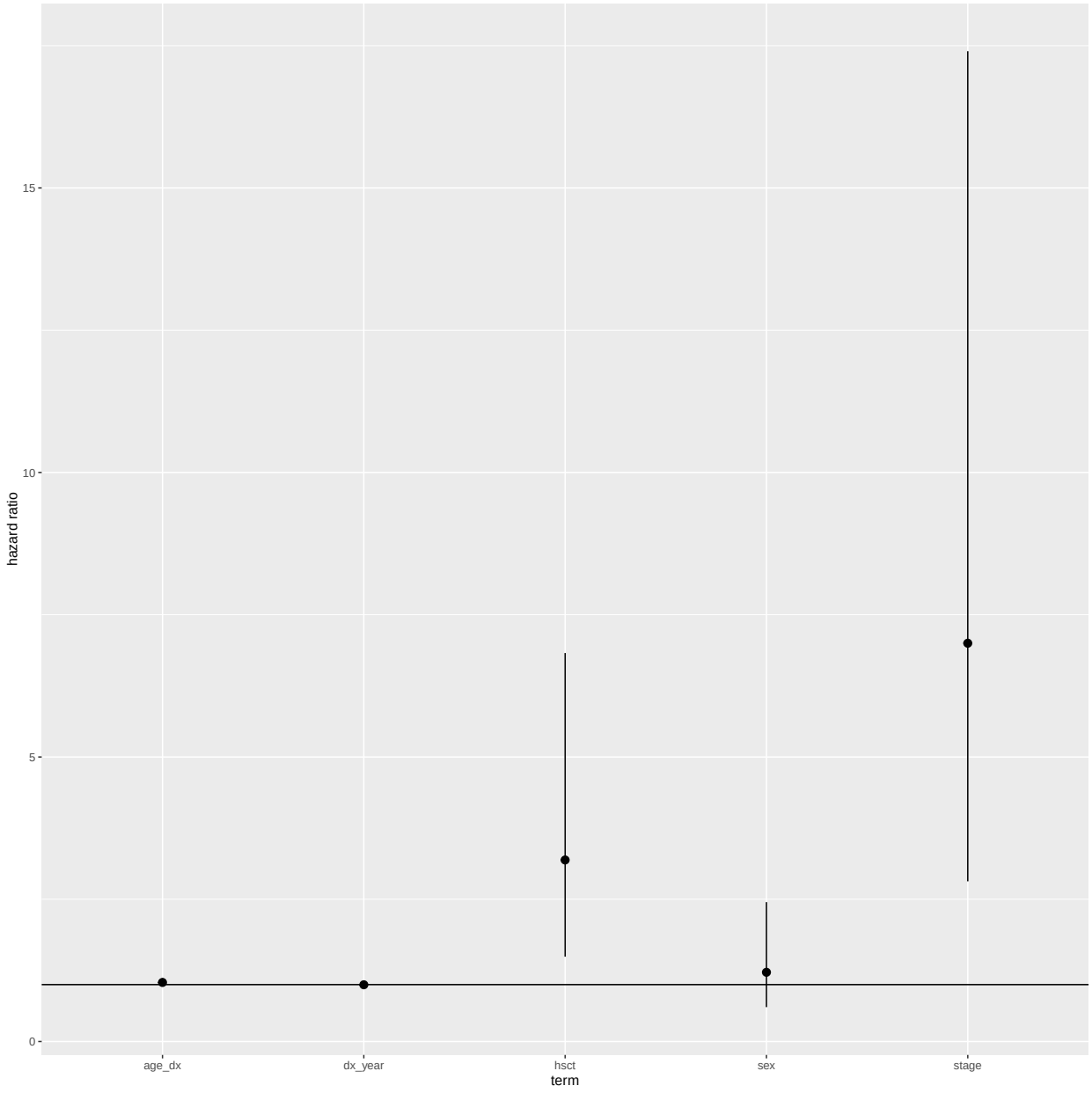
```
library(broom)
library(survival)

cox_dataset %>%
  coxph(formula = Surv(followup, death) ~ dx_year,
```

```
      data = .) %>%
  tidy() %>%
  ggplot() +
    geom_pointrange(aes(x = term,
                        y = exp(estimate),
                        ymin = exp(conf.low),
                        ymax = exp(conf.high))) +
    geom_hline(yintercept=1)
```



Let's adjust with age_dx, sex, stage, and hsct (which is confounded with factors associated with the treatment selection):

```r
library(dplyr)
library(broom)
library(survival)

cox_dataset %>%
  coxph(formula = Surv(followup, death) ~ age_dx + dx_year + sex + stage + hsct,
        data = .) %>%
  tidy() %>%
  ggplot() +
    geom_pointrange(aes(x = term,
                        y = exp(estimate),
                        ymin = exp(conf.low),
                        ymax = exp(conf.high))) +
    geom_hline(yintercept=1) +
  labs(y = "hazard ratio")
```



- There's basically no change: no association between diagnosis year and survival, given these variables.

10.3.3. Exploring the adjusted associations of stage and survival

The visualization of all of the models above show that point estimates of stage's hazard ratios vary between 3 and 15.5.

Let's see the unadjusted model:

```
library(dplyr)
library(broom)
library(survival)
```

```
cox_dataset %>%
  coxph(formula = Surv(followup, death) ~ stage,
        data = .) %>%
  tidy()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
stage	2.378439	0.4288454	5.546144	0	1.537917	3.21896

Getting stem cell transplantation treatment should be related to stage and other disease severity variables:

```
library(dplyr)
library(broom)
library(survival)

cox_dataset %>%
  coxph(formula = Surv(followup, death) ~ stage + hsct,
        data = .)
```

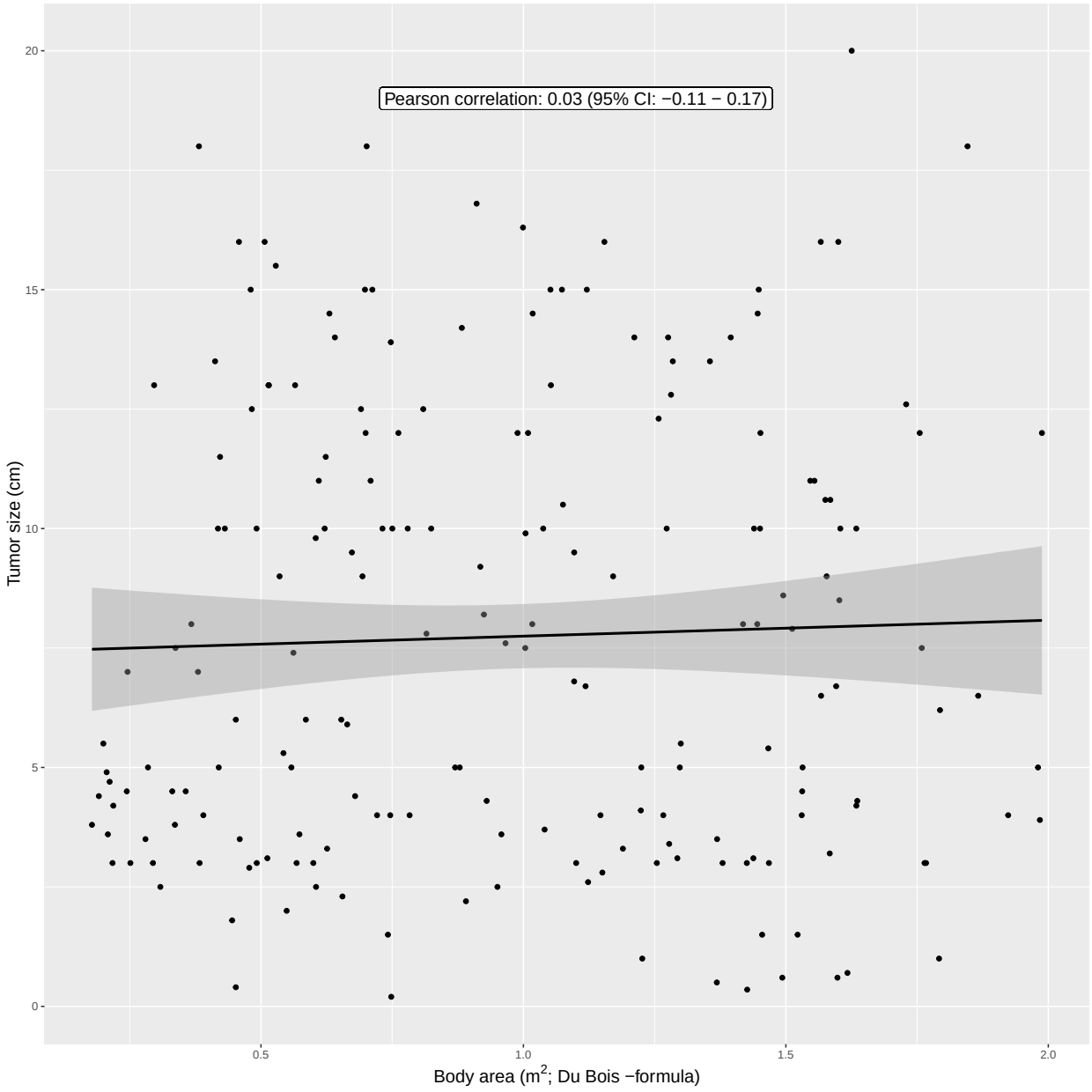
```
## Call:
## coxph(formula = Surv(followup, death) ~ stage + hsct, data = .)
##
##
##         coef exp(coef) se(coef)    z       p
## stage 1.946     7.003    0.462 4.21 2.5e-05
## hsct  1.137     3.117    0.385 2.95  0.0032
##
## Likelihood ratio test=48.6  on 2 df, p=2.8e-11
## n= 224, number of events= 32
##    (41 observations deleted due to missingness)
```

- Full model gave a point-estimate around HR = 6
- Unadjusted model gives a point-estimate around HR = 11
  - This is quite reasonable analysis, given the other variables, to get a measure of the predictive importance (at least) of stage.
- Adjusting with HSCT drops the point estimate to around HR = 7
  - The most likely explanation is that getting auto-HSCT -treatment reflects both stage and additional different disease severity factors far beyond the treatment effects (which are hard if not impossible to measure from this data).

11. Batch 8: Correlation between body area and tumor size

```
# Software
library(ggplot2)
library(Cairo)

# Dataset prepared in the Cox analysis -batch will be used
```

```r
areasize_data <- cox_dataset %>%
  select(tumorsize, area) %>%
  # Complete observations only
  na.omit()

# Correlation test
cor_area_size <- areasize_data %>%
  cor.test(~ area + tumorsize, data = .)

# Plot
areasize_data %>%
  ggplot() +
    geom_point(aes(x = area, y = tumorsize)) +
    geom_smooth(aes(x = area, y = tumorsize), method = "lm", color = "black") +
    geom_label(label = paste("Pearson correlation: ",
                             round(cor_area_size$estimate, 2),
                             " (95% CI: ",
                             round(cor_area_size$conf.int[1], 2),
                             " - ",
                             round(cor_area_size$conf.int[2], 2),
                             ")",
                             sep = ""),
               x = 1.1,
               y = 19,
               size = 5) +
    labs(y = "Tumor size (cm)",
         x = expression("Body area (" * m^2 * "; Du Bois -formula)", sep = "")) +
    theme(axis.title = element_text(size = 14))
```




```
ggsave("G:/Projects/potti/figures&tables/protocol-output/size-area-scatter.pdf")
```