Comparison of parametric and Random Forest MICE in imputation of missing data in survival analysis

Anoop D. Shah, Jonathan W. Bartlett, James Carpenter,
Owen Nicholas and Harry Hemingway

February 11, 2014

Contents

1 Introduction

This is a simulation study comparing various methods for imputation of missing covariate data in a survival analysis in which there are interactions between the predictor variables. We compare our new Random Forest method for MICE (Multivariate Imputation by Chained Equations) with other imputation methods and full data analysis. In our Random Forest method (RFcont), the conditional mean missing values are predicted using Random Forest and imputed values are drawn from Normal distributions centred on the predicted means [1].

We also perform a comparison with the methods recently published by Doove et al. [2]: mice.impute.cart (classification and regression trees in MICE) and mice.impute.rf (MICE using Random Forests).

2 Methods

We used the R packages **CALIBERrfimpute**, **survival**, **xtable**, **missForest** and **randomForest**. We created simulated survival datasets with two fully observed predictor variables (x_1, x_2) and a partially observed predictor (x_3), which depends on x_1, x_2 and their interaction. They were generated as follows:

x_1 Standard normal distribution

x_2 Standard normal distribution, independent of x_1

x_3 Derived from x_1 and x_2: $x_3 = 0.5(x_1 + x_2 - x_1.x_2) + e$ where e is normally distributed with mean 0 and variance 1.

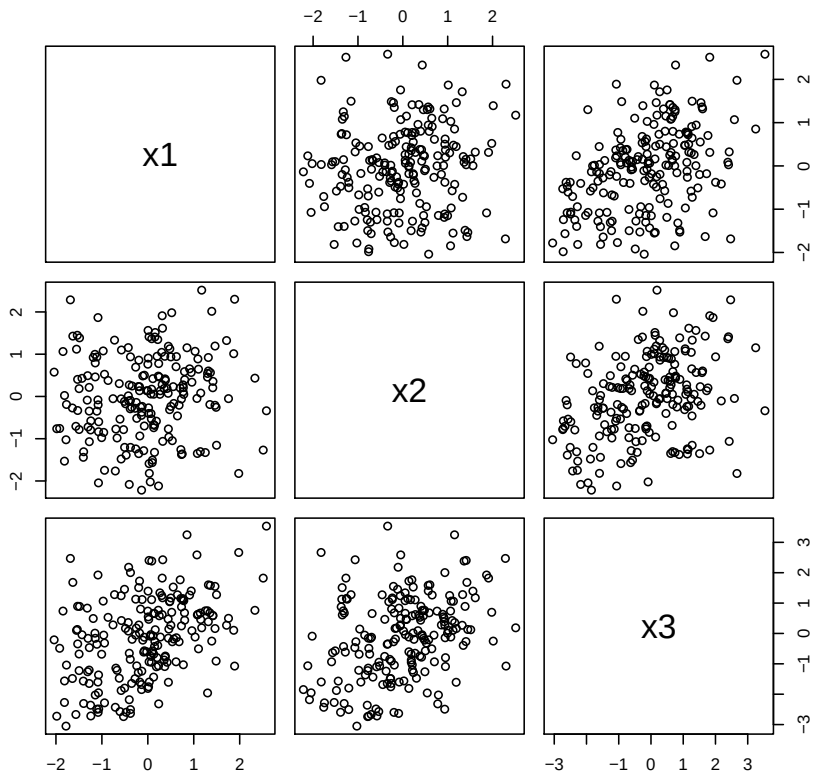
The equation for the log hazard of patient i was given by:

$$h_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \tag{1}$$

where all the β coefficients were set to 0.5.

We used an exponential distribution to generate a survival time for each patient. We also generated an observation time for each patient, as a random draw from a uniform distribution bounded by zero and the 50^{th} percentile of survival time. If the observation time was less than the survival time, the patient was considered as censored (event indicator 0, and the patient's follow-up ends on their censoring date), otherwise the event indicator was 1, with follow-up ending on the date of event.

Associations between predictor variables in a sample dataset



Linear regression model relating x_3 to x_1 and x_2:

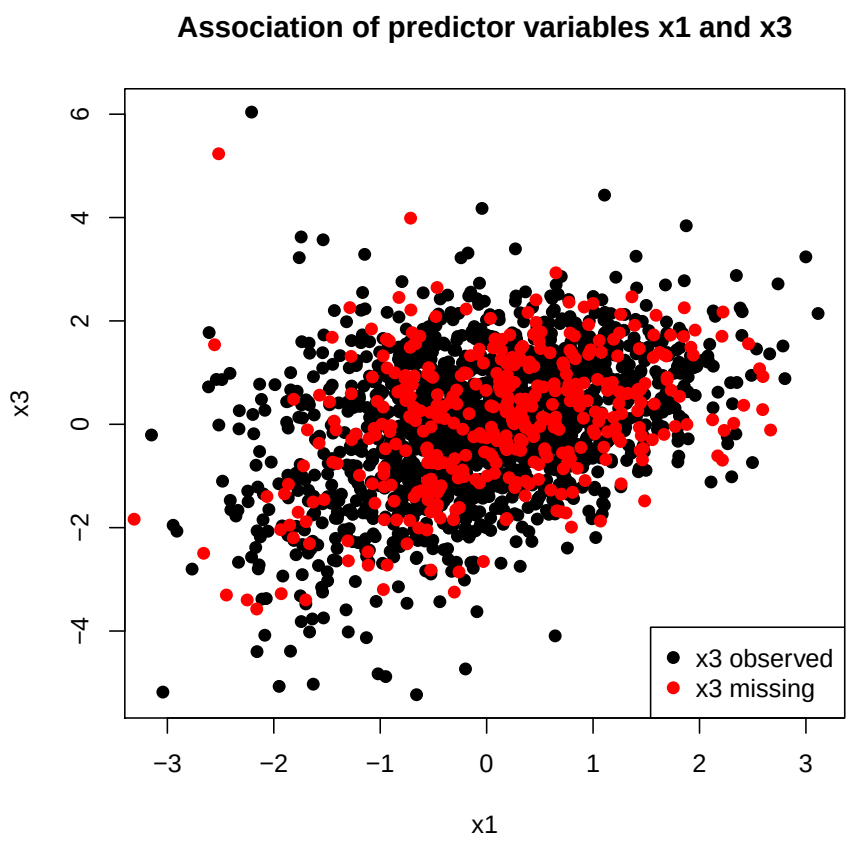
```
> summary(lm(x3 ~ x1*x2, data = mydata))

Call:
lm(formula = x3 ~ x1 * x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7876 -0.6788 -0.0065  0.6776  3.6570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.018379   0.007052  -2.606  0.00916 **
x1           0.491025   0.007005  70.096  < 2e-16 ***
x2           0.502028   0.007039  71.316  < 2e-16 ***
x1:x2       -0.498339   0.007065 -70.538  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.997 on 19996 degrees of freedom
Multiple R-squared:  0.4273,        Adjusted R-squared:  0.4272
F-statistic:  4974 on 3 and 19996 DF,  p-value: < 2.2e-16
```



All true log hazard ratios were assumed to be 0.5, with hazard ratios = 1.65. We checked that the hazard ratios in the simulated data were as expected for a large sample:

```
> # Cox proportional hazards analysis
> myformula <- as.formula(Surv(time, event) ~ x1 + x2 + x3)
```

```
> # Analysis with 10,000 simulated patients (or more
> # if the variable REFERENCE_SAMPLESIZE exists)
> if (!exists('REFERENCE_SAMPLESIZE')){
+         REFERENCE_SAMPLESIZE <- 10000
+ }
> # Use parallel processing, if available, to create
> # datasets more quickly.
> if ('parallel' %in% loadedNamespaces() &
+         !is.null(getOption('mc.cores')) &
+         Sys.info()['sysname'] == 'Linux'){
+         REFERENCE_SAMPLESIZE <- REFERENCE_SAMPLESIZE %/%
+                 getOption('mc.cores')
+         simdata <- mclapply(1:getOption('mc.cores'),
+                 function(x) makeSurv(REFERENCE_SAMPLESIZE))
+         simdata <- do.call('rbind', simdata)
+ } else {
+         simdata <- makeSurv(REFERENCE_SAMPLESIZE)
+ }
> summary(coxph(myformula, data = simdata))

Call:
coxph(formula = myformula, data = simdata)

  n= 100000, number of events= 30872

       coef exp(coef) se(coef)     z Pr(>|z|)
x1 0.498276  1.645882 0.006067 82.13   <2e-16 ***
x2 0.515256  1.674067 0.006097 84.51   <2e-16 ***
x3 0.498885  1.646885 0.005315 93.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   exp(coef) exp(-coef) lower .95 upper .95
x1     1.646     0.6076     1.626     1.666
x2     1.674     0.5973     1.654     1.694
x3     1.647     0.6072     1.630     1.664

Concordance= 0.762  (se = 0.002 )
Rsquare= 0.251   (max possible= 0.999 )
Likelihood ratio test= 28947  on 3 df,   p=0
Wald test            = 24646  on 3 df,   p=0
Score (logrank) test = 24437  on 3 df,   p=0
```

We created datasets containing 2000 simulated patients. For each dataset, we first analysed the complete dataset with no values missing, then artificially created missingness in variable x_3, imputed the missing values using various methods, and analysed the imputed datasets. We combined parameter estimates from multiply imputed datasets using Rubin's rules.

2.1 Missingness mechanism

Missingness was imposed in x_3 dependent on x_1, x_2, the event indicator and the marginal Nelson-Aalen cumulative hazard, using a logistic regression model. The linear predictors

were offset by an amount chosen to make the overall proportion of each variable missing approximately 0.2, i.e.:

$$P(\text{miss})_i = \frac{\exp(lp_i + \text{offset})}{1 + \exp(lp_i + \text{offset})} \tag{2}$$

$$lp_i = 0.1x_{1i} + 0.1x_{2i} + 0.1 \times \text{cumhaz}_i + 0.1 \times \text{event}_i \tag{3}$$

where 'event' is the event indicator and 'cumhaz' is the marginal Nelson-Aalen cumulative hazard.

We analysed the datasets with missing data using different methods of multiple imputation. We calculated the marginal Nelson-Aalen cumulative hazard and included it in all imputation models, along with the event indicator and follow-up time.

```
> # Setting analysis parameters: To analyse more than 3 samples,
> # set N to the desired number before running this program
> if (!exists('N')){
+         N <- 3
+ }
> # Number of imputations (set to at least 10 when
> # running an actual simulation)
> if (!exists('NIMPS')){
+         NIMPS <- 3
+ }
> # Use parallel processing if the 'parallel' package is loaded
> if ('parallel' %in% loadedNamespaces() & Sys.info()['sysname'] == 'Linux'){
+         cat('Using parallel processing\n')
+         results <- mclapply(1:N, doanalysis)
+ } else {
+         results <- lapply(1:N, doanalysis)
+ }

Using parallel processing
```

We used the following methods of multiple imputation. The number of imputations was 10. In each case, the imputation model for x_3 contained x_1, x_2, the event indicator and the marginal Nelson-Aalen cumulative hazard:

rfImpute – a single imputation method in the randomForest package, which uses the proximity matrix of a Random Forest to impute missing predictor variables. The marginal Nelson-Aalen cumulative hazard was considered as the 'response' for the purpose of running this algorithm. It was run with 300 trees (default) and 10 iterations.

missForest – from the missForest package, which completes a dataset in an iterative way using Random Forest prediction. It was run with maximum 10 iterations (default) and 100 trees per forest (default).

CART MICE – Clasification and regression tree MICE method from the mice package (mice.impute.cart).

RF MICE (Doove) – Random Forest MICE method from Doove et al. [2], which is available as function mice.impute.rf in the mice package.

RFcont MICE – Random Forest MICE method for continuous variables from the CALIBERrfimpute package with 5, 10, 20, 50 or 100 trees.

Parametric MICE – normal-based linear regression with default settings, in which the imputation model for x_3 is of the form:

$$x_3 = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \beta_3.\text{event} + \beta_4.\text{cumhaz} + e$$

where e is the residual variance.

We analysed 1000 samples. We calculated the following for each method and each parameter:

- Bias of log hazard ratio
- Standard error of bias (Monte Carlo error)
- Z-score for bias
- Standard deviation of estimated log hazard ratio
- Mean length of 95% confidence intervals
- Coverage of 95% confidence intervals (proportion containing the true log hazard ratio)

3 Results

All the true log hazard ratios were set at 0.5. The Z-score is defined as the mean bias divided by the empirical standard error of the estimates.

3.1 Fully observed variables

Log hazard ratio for the continuous fully observed variable x_1:

	Bias	Standard error of bias	Z-score for bias	SD of estimate	Mean 95% CI length	95% CI coverage
Full data	-0.00213	0.00131	-1.63	0.0413	0.169	0.957
rfImpute	-0.0268	0.00136	-19.7	0.043	0.169	0.908
missForest	-0.0219	0.00138	-15.9	0.0436	0.169	0.926
CART MICE	-0.00734	0.00137	-5.36	0.0433	0.172	0.954
RF MICE (Doove)	-0.000256	0.00132	-0.194	0.0418	0.173	0.962
RFcont MICE, 5 trees	-0.00603	0.00133	-4.53	0.0421	0.175	0.965
RFcont MICE, 10 trees	-0.00868	0.00134	-6.5	0.0423	0.174	0.96
RFcont MICE, 20 trees	-0.0105	0.00134	-7.84	0.0424	0.173	0.959
RFcont MICE, 50 trees	-0.0113	0.00134	-8.38	0.0425	0.173	0.957
RFcont MICE, 100 trees	-0.0117	0.00134	-8.77	0.0424	0.173	0.957
Parametric MICE	-0.0346	0.00141	-24.6	0.0446	0.176	0.897

Log hazard ratio for the continuous fully observed variable x_2:

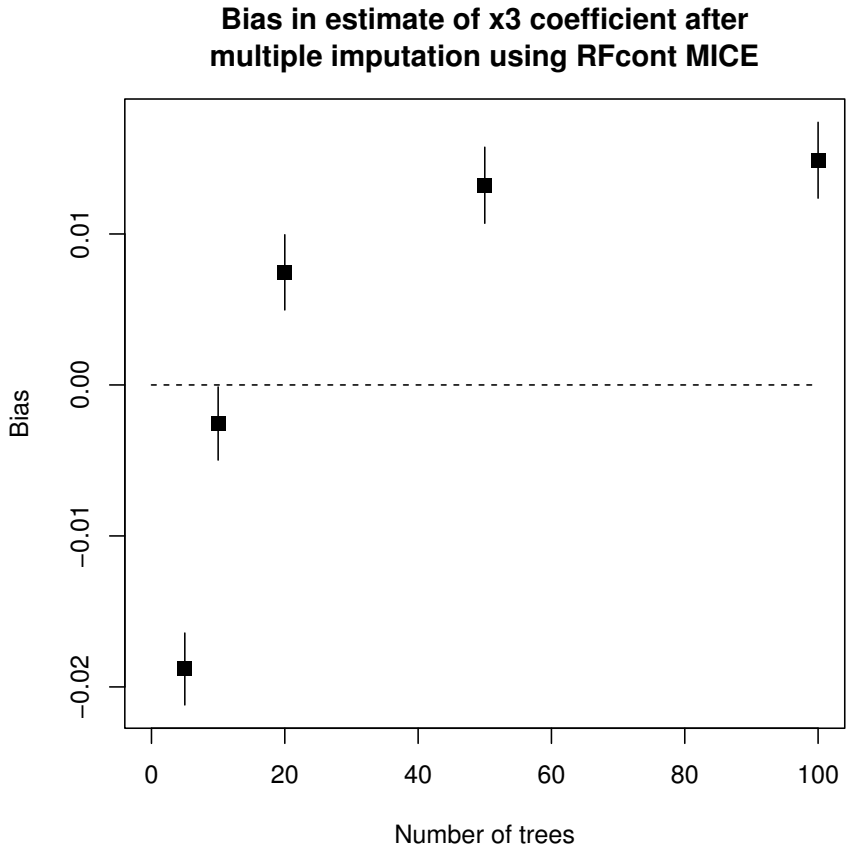
	Bias	Standard error of bias	Z-score for bias	SD of estimate	Mean 95% CI length	95% CI coverage
Full data	-0.000581	0.00139	-0.419	0.0438	0.169	0.953
rfImpute	-0.0248	0.00142	-17.5	0.0449	0.169	0.898
missForest	-0.0201	0.00145	-13.9	0.0458	0.169	0.907
CART MICE	-0.00579	0.00143	-4.03	0.0454	0.172	0.948
RF MICE (Doove)	0.00227	0.00138	1.64	0.0438	0.173	0.949
RFcont MICE, 5 trees	-0.00358	0.0014	-2.56	0.0444	0.175	0.954
RFcont MICE, 10 trees	-0.00662	0.0014	-4.72	0.0444	0.174	0.951
RFcont MICE, 20 trees	-0.00826	0.00141	-5.86	0.0445	0.174	0.951
RFcont MICE, 50 trees	-0.00897	0.00141	-6.38	0.0445	0.173	0.948
RFcont MICE, 100 trees	-0.00948	0.00141	-6.72	0.0446	0.173	0.947
Parametric MICE	-0.0323	0.00148	-21.9	0.0467	0.176	0.877

3.2 Partially observed variable

Log hazard ratio for the continuous partially observed variable x_3:

	Bias	Standard error of bias	Z-score for bias	SD of estimate	Mean 95% CI length	95% CI coverage
Full data	0.00155	0.00118	1.31	0.0374	0.148	0.943
rfImpute	0.12	0.00138	86.4	0.0438	0.154	0.168
missForest	0.0793	0.00146	54.3	0.0462	0.158	0.514
CART MICE	-0.00669	0.00137	-4.87	0.0434	0.16	0.929
RF MICE (Doove)	-0.00555	0.00121	-4.6	0.0382	0.166	0.972
RFcont MICE, 5 trees	-0.0188	0.00122	-15.5	0.0384	0.171	0.958
RFcont MICE, 10 trees	-0.00255	0.00124	-2.07	0.0391	0.168	0.97
RFcont MICE, 20 trees	0.00746	0.00127	5.87	0.0402	0.167	0.954
RFcont MICE, 50 trees	0.0132	0.00129	10.3	0.0407	0.166	0.945
RFcont MICE, 100 trees	0.0149	0.00128	11.6	0.0405	0.166	0.941
Parametric MICE	-0.0519	0.00124	-41.7	0.0393	0.178	0.815

The following graph shows the bias for RFcont MICE methods by number of trees (bias estimated from 1000 simulations; the lines denote 95% confidence intervals):



3.3 Pairwise comparisons between methods

3.3.1 Comparison of bias

Difference between absolute bias (negative means that the first method is less biased). P values from paired sample t tests. Significance level: * P <0.05, ** P <0.01, *** P <0.001.

Coefficient	RFcont MICE 10 vs parametric MICE	RFcont MICE 100 vs parametric MICE	RFcont MICE 10 vs RFcont MICE 100
x1	-0.0259 ***	-0.0229 ***	-0.00306 ***
x2	-0.0257 ***	-0.0228 ***	-0.00286 ***
x3	-0.0493 ***	-0.037 ***	-0.0123 ***

Coefficient	RFcont MICE 10 vs RF MICE Doove	RFcont MICE 10 vs CART MICE	RF MICE Doove vs CART MICE
x1	0.00843 ***	0.00135 ***	-0.00708 ***
x2	0.00435	0.000834 **	-0.00352
x3	-0.003 ***	-0.00413 ***	-0.00113 *

3.3.2 Comparison of precision

Ratio of variance of estimates (less than 1 means that the first method is more precise). P values from F test. Significance level: * P <0.05, ** P <0.01, *** P <0.001.

Coefficient	RFcont MICE 10 vs parametric MICE	RFcont MICE 100 vs parametric MICE	RFcont MICE 10 vs RFcont MICE 100
x1	0.898	0.903	0.995
x2	0.904	0.913	0.99
x3	0.987	1.06	0.929

Coefficient	RFcont MICE 10 vs RF MICE (Doove)	RFcont MICE 10 vs CART MICE	RF MICE (Doove) vs CART MICE
x1	1.02	0.953	0.932
x2	1.03	0.956	0.931
x3	1.05	0.81 ***	0.773 ***

3.3.3 Comparison of confidence interval length

Ratio of mean length of 95% confidence intervals (less than 1 means that the first method produces smaller confidence intervals). P values from paired sample t test. Significance level: * P <0.05, ** P <0.01, *** P <0.001.

Coefficient	RFcont MICE 10 vs parametric MICE	RFcont MICE 100 vs parametric MICE	RFcont MICE 10 vs RFcont MICE 100
x1	0.9862 ***	0.9806 ***	1.006 ***
x2	0.9864 ***	0.981 ***	1.006 ***
x3	0.9453 ***	0.934 ***	1.012 ***

Coefficient	RFcont MICE 10 vs RF MICE (Doove)	RFcont MICE 10 vs CART MICE	RF MICE (Doove) vs CART MICE
x1	1.004 ***	1.013 ***	1.009 ***
x2	1.003 ***	1.012 ***	1.009 ***
x3	1.014 ***	1.048 ***	1.033 ***

3.3.4 Comparison of confidence interval coverage

Difference between percentage coverage of 95% confidence intervals (positive means that the first method has greater coverage). P values for pairwise comparisons by McNemar's test. Significance level: * P <0.05, ** P <0.01, *** P <0.001.

Coefficient	RFcont MICE 10 vs parametric MICE	RFcont MICE 100 vs parametric MICE	RFcont MICE 10 vs RFcont MICE 100
x1	6.3 ***	6.0 ***	0.3
x2	7.4 ***	7.0 ***	0.4
x3	15.5 ***	12.6 ***	2.9 ***

Coefficient	RFcont MICE 10 vs RF MICE (Doove)	RFcont MICE 10 vs CART MICE	RF MICE (Doove) vs CART MICE
x1	-0.2	0.6	0.8
x2	0.2	0.3	0.1
x3	-0.2	4.1 ***	4.3 ***

4 Discussion

In this simulation, parametric MICE using the default settings yielded a biased estimate for the coefficient for the partially observed variable x_3. This is because the interaction between x_1 and x_2 was not included in the imputation models. The estimate using the CART or Random Forest MICE methods were less biased, more precise and had shorter confidence intervals with greater coverage. Omissions of interactions between predictors can potentially result in bias using parametric MICE even if, as in this case, the interaction is not present in the substantive model.

4.1 CART versus Random Forest MICE

CART MICE produced estimates for the x_3 coefficient that were less precise than the Random Forest MICE methods, and coverage of 95% confidence intervals was only 93%.

4.2 Comparison of Random Forest MICE methods

Coefficients estimated after imputation using CART or Random Forest MICE methods were slightly biased. The bias was statistically significant but small in magnitude. Using RFcont MICE, the x_3 coefficient was biased towards the null with 5 or 10 trees and biased away from the null with 20 or more trees; bias was minimised using 10 or 20 trees.

Confidence intervals estimated using Doove's Random Forest MICE method were slightly shorter than those obtained using RFcont MICE but coverage was >95% with both methods.

Computation time for Random Forest MICE methods was proportional to the number of trees; Doove's method with 100 trees took about 10 times as long as RFcont with 10 trees.

4.3 rfImpute and missForest

Parameters estimated after imputation using rfImpute and missForest were biased and the coverage of 95% confidence intervals was less than 95%. These methods are not suitable for multiple imputation. Failure to draw from the correct conditional distribution leads to bias and underestimation of the uncertainty when statistical models are fitted to imputed data.

4.4 Implications for further research

This simulation demonstrates a situation in which Random Forest MICE methods have an advantage over parametric MICE. Both Doove's method (RF) and our method (RFcont) performed well, and on some performance measures Doove's method was superior.

It would be useful to compare these methods in simulations based on real datasets.

5 Appendix: R code

5.1 R functions

This R code needs to be run in order to load the necessary functions before running the script (Section 5.2).

5.1.1 Data generating functions

```
makeSurv <- function(n = 2000, loghr = kLogHR){
        # Creates a survival cohort of n patients. Assumes that censoring is
        # independent of all other variables

        # x1 and x2 are random normal variables
        data <- data.frame(x1 = rnorm(n), x2 = rnorm(n))

        # Create the x3 variable
        data$x3 <- 0.5 * (data$x1 + data$x2 - data$x1 * data$x2) + rnorm(n)

        # Underlying log hazard ratio for all variables is the same
        data$y <- with(data, loghr * (x1 + x2 + x3))
        data$survtime <- rexp(n, exp(data$y))

        # Censoring - assume uniform distribution of observation times
        # up to a maximum
        obstime <- runif(nrow(data), min = 0,
                max = quantile(data$survtime, 0.5))
        data$event <- as.integer(data$survtime <= obstime)
        data$time <- pmin(data$survtime, obstime)
        data$cumhaz <- nelsonaalen(data, time, event)

        # True log hazard and survival time are not seen in the data
        # so remove them
        data$y <- NULL
        data$survtime <- NULL

        return(data)
}

makeMarSurv <- function(data, pmissing = kPmiss){
        # Introduces missing data dependent on event indicator
        # and cumulative hazard and x1 and x2

        logistic <- function(x){
                exp(x) / (1 + exp(x))
        }

        predictions <- function(lp, n){
                # uses the vector of linear predictions (lp) from a logistic model
                # and the expected number of positive responses (n) to generate
                # a set of predictions by modifying the baseline

                trialn <- function(lptrial){
                        sum(logistic(lptrial))
```

```
                }
                stepsize <- 32
                lptrial <- lp
                while(abs(trialn(lptrial) - n) > 1){
                        if (trialn(lptrial) > n){
                                # trialn bigger than required
                                lptrial <- lptrial - stepsize
                        } else {
                                lptrial <- lptrial + stepsize
                        }
                        stepsize <- stepsize / 2
                }
                # Generate predictions from binomial distribution
                as.logical(rbinom(logical(length(lp)), 1, logistic(lptrial)))
        }

        data$x3[predictions(0.1 * data$x1 + 0.1 * data$x2 +
                0.1 * data$cumhaz + 0.1 * data$event, nrow(data) * pmissing)] <- NA
        return(data)
}
```

5.1.2 Functions to analyse data

```
coxfull <- function(data){
        # Full data analysis
        coefs <- summary(coxph(myformula, data = data))$coef
        # return a vector of coefficients (est), upper and lower 95% limits
        confint <- cbind(coefs[, 'coef'] - qnorm(0.975) * coefs[, 'se(coef)'],
                coefs[, 'coef'] + qnorm(0.975) * coefs[, 'se(coef)'])
        out <- cbind(coefs[, 'coef'], confint,
                kLogHR >= confint[,1] & kLogHR <= confint[,2])
        colnames(out) <- c('est', 'lo 95', 'hi 95', 'cover')
        out
}

coximpute <- function(imputed_datasets){
        # Analyses a list of imputed datasets
        docoxmodel <- function(data){
                coxph(myformula, data=data)
        }
        mirafits <- as.mira(lapply(imputed_datasets, docoxmodel))
        out <- summary(pool(mirafits))
        out <- cbind(out, kLogHR >= out[, 'lo 95'] & kLogHR <= out[, 'hi 95'])
        # Whether this confidence interval contains the true hazard ratio
        colnames(out)[length(colnames(out))] <- 'cover'
        out
}

domissf <- function(missdata, reps = NIMPS){
        # Imputation by missForest
        out <- list()
        for (i in 1:reps){
                invisible(capture.output(
```

```
                        out[[i]] <- missForest(missdata)$ximp))
        }
        out
}

dorfimpute <- function(missdata, reps = NIMPS){
        # Cumulative hazard is the outcome, filling in missing data
        # using proximities, using a 300-tree Random Forest.
        out <- list()
        for (i in 1:reps){
                invisible(capture.output(
                        out[[i]] <- rfImpute(cumhaz ~ ., missdata, iter = 10)))
        }
        out
}

mice.impute.cart <- function(y, ry, x, minbucket = 5, cp = 1e-04,
        ...){
        xobs <- x[ry,]
        xmis <- x[!ry,]
        yobs <- y[ry]
        if (is.factor(yobs)==F){
        fit <- rpart(yobs~., data = cbind(yobs,xobs), method = "anova",
        control = rpart.control(minbucket = minbucket, cp = cp), ...)
                leafnr  <- floor(as.numeric(row.names(fit$frame[fit$where,])))
                fit$frame$yval <- as.numeric(row.names(fit$frame))
                nodes <- predict(object = fit, newdata = xmis)
                donor <- lapply(nodes, function(s) yobs[leafnr == s])
                impute <- sapply(1:length(donor), function(s){
                        sample(donor[[s]], 1)
                })
        } else {
                fit <- rpart(yobs~., data = cbind(yobs, xobs),
                method = "class", control = rpart.control(
                minbucket = minbucket, cp = cp), ...)
                nodes <- predict(object = fit, newdata = xmis)
                impute <- apply(nodes, MARGIN = 1, FUN = function(s){
                        sample(colnames(nodes), size = 1, prob = s)
                })
        }
        return(impute)
}

mice.impute.rf <- function(y, ry, x, ntrees = 100,
        nodesize = 5, ...){
        # Use default mtry, i.e. one third the number of predictors for
        # categorical variables, square root of the number of predictors
        # for continuous dependent variables
        xobs <- x[ry,]
        xmis <- x[!ry,]
        yobs <- y[ry]
        # Function to create a single tree
        onetree <- function(xobs, xmis, yobs){
```

```
                fit <- randomForest(yobs ~ ., data = cbind(yobs, xobs),
                        ntree = 1, replace = TRUE, type = regression,
                        sampsize = length(yobs), nodesize = 5)
                leafnr <- predict(object = fit, newdata = xobs,
                        nodes = T)
                nodes <- predict(object = fit, newdata = xmis,
                        nodes = T)
                # Return a vector of observed y values that are
                # part of the same terminal leaf as the predicted
                # missing y value
                donor <- lapply(nodes, function(s){
                        yobs[leafnr == s]
                })
                return(donor)
        }
        # Create a matrix of vectors of donors, from the number
        # of trees desired
        forest <- sapply(1:(ntrees = ntrees), FUN = function(s){
                onetree(xobs, xmis, yobs)
        })
        # For each missing value, randomly choose a donor value
        # from any of the possible donor values across all trees
        impute <- apply(forest, MARGIN = 1, FUN = function(s){
                sample(unlist(s), 1)
        })
        return(impute)
}

mice.impute.rfcont5 <- function(y, ry, x, ...){
        mice.impute.rfcont(y = y, ry = ry, x = x, ntree_cont = 5)
}

mice.impute.rfcont10 <- function(y, ry, x, ...){
        mice.impute.rfcont(y = y, ry = ry, x = x, ntree_cont = 10)
}

mice.impute.rfcont20 <- function(y, ry, x, ...){
        mice.impute.rfcont(y = y, ry = ry, x = x, ntree_cont = 20)
}

mice.impute.rfcont50 <- function(y, ry, x, ...){
        mice.impute.rfcont(y = y, ry = ry, x = x, ntree_cont = 50)
}

mice.impute.rfcont100 <- function(y, ry, x, ...){
        mice.impute.rfcont(y = y, ry = ry, x = x, ntree_cont = 100)
}

domice <- function(missdata, functions, reps = NIMPS){
        mids <- mice(missdata, defaultMethod = functions,
                m = reps, visitSequence = 'monotone',
                printFlag = FALSE, maxit = 10)
        lapply(1:reps, function(x) complete(mids, x))
}
```

```
doanalysis <- function(x){
        # Creates a dataset, analyses it using different methods, and outputs
        # the result as a matrix of coefficients / SE and coverage
        data <- makeSurv(kSampleSize)
        missdata <- makeMarSurv(data)
        out <- list()
        out$full <- coxfull(data)
        out$rfimpute <- coximpute(dorfimpute(missdata))
        out$missf <- coximpute(domissf(missdata))
        out$rf5 <- coximpute(domice(missdata, 'rfcont5'))
        out$rf10 <- coximpute(domice(missdata, 'rfcont10'))
        out$rf20 <- coximpute(domice(missdata, 'rfcont20'))
        out$rf50 <- coximpute(domice(missdata, 'rfcont50'))
        out$rf100 <- coximpute(domice(missdata, 'rfcont100'))
        out$rf <- coximpute(domice(missdata, 'rf'))
        out$cart <- coximpute(domice(missdata, 'cart'))
        out$mice <- coximpute(domice(missdata, 'norm'))
        out
}
```

5.1.3 Functions to compare methods

```
pstar <- function(x){
        if (x < 0.001){
                '***'
        } else if (x < 0.01){
                '**'
        } else if (x < 0.05){
                '*'
        } else {
                ''
        }
}

compareBias <- function(method1, method2){
        # Generates a table comparing bias
        # Comparison statistic is the difference in absolute bias
        # (negative means first method is better)

        compareBiasVar <- function(varname){
                # All coefficients should be kLogHR
                bias1 <- sapply(results, function(x){
                        x[[method1]][varname, 'est']
                }) - kLogHR
                bias2 <- sapply(results, function(x){
                        x[[method2]][varname, 'est']
                }) - kLogHR

                if (sign(mean(bias1)) == -1){
                        bias1 <- -bias1
                }
                if (sign(mean(bias2)) == -1){
                        bias2 <- -bias2
```

```
                }

                paste(formatC(mean(bias1) - mean(bias2), format = 'fg', digits = 3),
                        pstar(t.test(bias1 - bias2)$p.value))
        }

        sapply(variables, compareBiasVar)
}

compareVariance <- function(method1, method2){
        # Generates a table comparing precision between two methods
        # Comparison statistic is ratio of variance
        # (smaller means first method is better)

        compareVarianceVar <- function(varname){
                e1 <- sapply(results, function(x){
                        x[[method1]][varname, 'est']
                })
                e2 <- sapply(results, function(x){
                        x[[method2]][varname, 'est']
                })
                paste(formatC(var(e1) / var(e2), format = 'fg', digits = 3),
                        pstar(var.test(e1, e2)$p.value))
        }

        sapply(variables, compareVarianceVar)
}

compareCIlength <- function(method1, method2){
        # Generates a table comparing coverage percentage between two methods
        # Comparison statistic is the ratio of confidence interval lengths
        # (less than 1 = first better)

        compareCIlengthVar <- function(varname){
                # Paired t test for bias (difference in estimate)
                len1 <- sapply(results, function(x){
                        x[[method1]][varname, 'hi 95'] -
                                x[[method1]][varname, 'lo 95']
                })
                len2 <- sapply(results, function(x){
                        x[[method2]][varname, 'hi 95'] -
                                x[[method2]][varname, 'lo 95']
                })

                paste(formatC(mean(len1) / mean(len2), format = 'fg', digits = 3),
                        pstar(t.test(len1 - len2)$p.value))
        }

        sapply(variables, compareCIlengthVar)
}

compareCoverage <- function(method1, method2){
        # Generates a table comparing coverage percentage between two methods
```

```
        # Comparison statistic is the difference in coverage
        # (positive = first better)

        compareCoverageVar <- function(varname){
                # Paired t test for bias (difference in estimate)

                cov1 <- sapply(results, function(x){
                        x[[method1]][varname, 'cover']
                })
                cov2 <- sapply(results, function(x){
                        x[[method2]][varname, 'cover']
                })

                paste(formatC(100 * (mean(cov1) - mean(cov2)), format = 'f',
                        digits = 1),
                        pstar(binom.test(c(sum(cov1 == TRUE  & cov2 == FALSE),
                        sum(cov1 == FALSE & cov2 == TRUE)))$p.value))
        }

        sapply(variables, compareCoverageVar)
}
```

5.1.4 Functions to compile and display results

```
getParams <- function(coef, method){
        estimates <- sapply(results, function(x){
                x[[method]][coef, 'est']
        })
        bias <- mean(estimates) - kLogHR
        se_bias <- sd(estimates) / sqrt(length(estimates))
        z <- bias / se_bias
        ci_len <- mean(sapply(results, function(x){
                x[[method]][coef, 'hi 95'] - x[[method]][coef, 'lo 95']
        }))
        ci_cov <- mean(sapply(results, function(x){
                x[[method]][coef, 'cover']
        }))
        out <- c(bias, se_bias, z, sd(estimates), ci_len, ci_cov)
        names(out) <- c('bias', 'se_bias', 'z_bias', 'sd', 'ci_len', 'ci_cov')
        out
}

showTable <- function(coef){
        methods <- c('full', 'rfimpute', 'missf', 'cart', 'rf',
                'rf5', 'rf10', 'rf20', 'rf50', 'rf100', 'mice')
        methodnames <- c('Full data', 'rfImpute', 'missForest',
                'CART MICE', 'RF MICE (Doove)',
                paste('RFcont MICE, ', c(5, 10, 20, 50, 100), 'trees'),
                'Parametric MICE')
        out <- t(sapply(methods, function(x){
                getParams(coef, x)
        }))
        out <- formatC(out, digits = 3, format = 'fg')
```

```
        out <- rbind(c('', 'Standard', 'Z-score ', 'SD of', 'Mean 95%', '95% CI'),
                c('Bias', 'error of bias', 'for bias', 'estimate',
                'CI length', 'coverage'), out)
        out <- cbind(c('', '', methodnames), out)
        print(xtable(out), floating = FALSE, include.rownames = FALSE,
                include.colnames = FALSE, hline.after = c(0, 2, nrow(out)))
}

maketable <- function(comparison){
        # comparison is a function such as compareCoverage, compareBias
        compare <- cbind(comparison('rf10', 'mice'),
                comparison('rf100', 'mice'),
                comparison('rf10', 'rf100'),
                comparison('rf10', 'rf'))
        compare <- cbind(rownames(compare), compare)
        compare <- rbind(
                c('', 'RF MICE 10 vs', 'RF MICE 100 vs',
                        'RF MICE 10 vs', 'RF MICE 10 vs'),
                c('Coefficient', 'parametric MICE',
                        'parametric MICE', 'RF MICE 100',
                        'RF MICE (Doove)'),
                compare)
        print(xtable(compare), include.rownames = FALSE,
                include.colnames = FALSE, floating = FALSE,
                hline.after = c(0, 2, nrow(compare)))
}
```

5.2 R script

Run this script after loading the functions above.

```
# Install CALIBERrfimpute if necessary:
# install.packages("CALIBERrfimpute", repos="http://R-Forge.R-project.org")
library(CALIBERrfimpute)
library(missForest)
library(survival)
library(xtable)
library(parallel) # Use parallel processing on Unix
library(rpart)    # Recursive partitioning (trees)

# Initialise constants
kPmiss <- 0.2 # probability of missingness
kLogHR <- 0.5 # true log hazard ratio
kSampleSize <- 2000 # number of patients in simulated datasets

# Set number of samples
N <- 1000

# Perform the simulation
results <- mclapply(1:N, doanalysis)

# Show results
showTable('x1'); showTable('x2'); showTable('x3')
```

```
# Names of the variables in the comparison
variables <- c('x1', 'x2', 'x3')

# Show comparisons between methods
maketable(compareBias)
maketable(compareVariance)
maketable(compareCIlength)
maketable(compareCoverage)
```

References

[1] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology* 2014. doi: 10.1093/aje/kwt312

[2] Doove LL, van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis* 2014;72:92–104. doi: 10.1016/j.csda.2013.10.025