Missing Data Analyses
FAQ

Ehsan Karim
ehsan.karim@ubc.ca
SPPH 604 Discussions

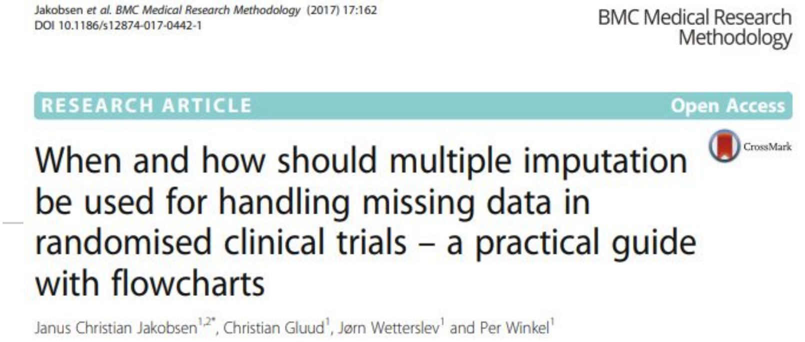
Missingness Assumptions

MAR vs. MCAR: MCAR missingness doesn't follow any pattern. From empirical data, we may be able to disprove this (reject null hypothesis of MCAR if there is a pattern).

While it may be possible to reject MCAR (meaning either MAR or MNAR is more likely), it is not possible to say which one is more likely (MAR or MNAR) just based on data analysis.

Amaranth purple

Complete Case


Jakobsen *et al. BMC Medical Research Methodology* (2017) 17:162
DOI 10.1186/s12874-017-0442-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts

Janus Christian Jakobsen[1,2*], Christian Gluud[1], Jørn Wetterslev[1] and Per Winkel[1]

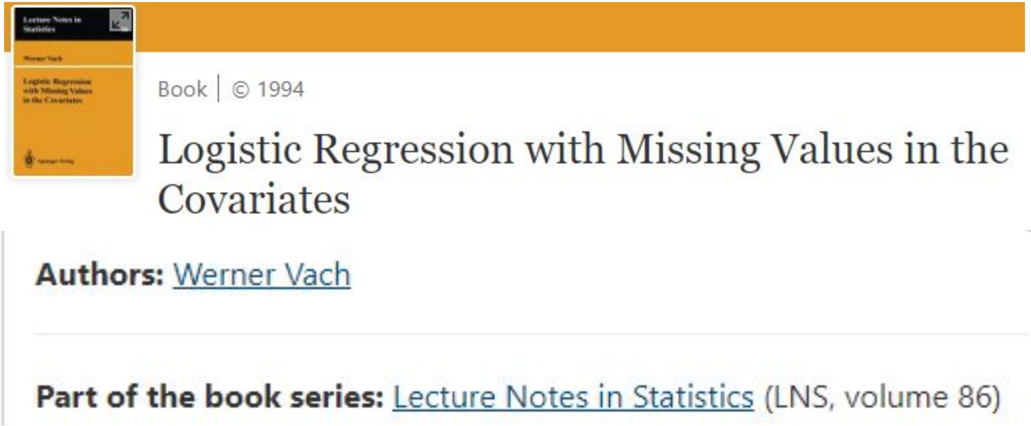
Rule of thumb:

Complete case (CC) analysis could be used as the primary analysis if

- % of missing observations (for all variables combined) are below ~5%
- When potential impacts of the presence of missing data is negligible
- Best-worst and worst-best case sensitivity analyses could be used as a sensitivity analysis
  - SES = 1 for all missing vs. SES = 5 for all)
- Only outcome variable (of primary analysis) has missing, CC will be more efficient than MI.
- If relatively certain that the data are MCAR (don't base your decision solely on Little's test)

https://ehsanx.github.io/EpiMethods/missingdata6.html

Ad hoc methods



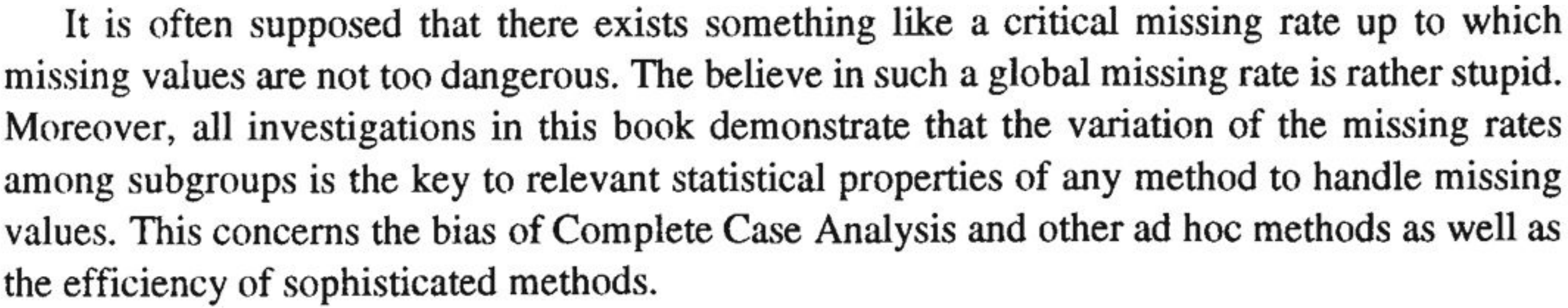
Book | © 1994

Logistic Regression with Missing Values in the Covariates

Authors: Werner Vach

Part of the book series: Lecture Notes in Statistics (LNS, volume 86)

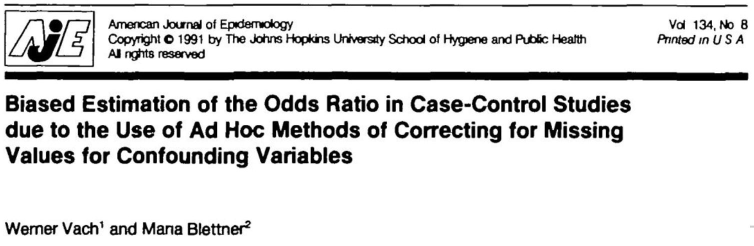
Harsh words from methodologists; so could be the thought process of the reviewer! Hence, if using an ad hoc method, should have a very clear justification!!



It is often supposed that there exists something like a critical missing rate up to which missing values are not too dangerous. The believe in such a global missing rate is rather stupid. Moreover, all investigations in this book demonstrate that the variation of the missing rates among subgroups is the key to relevant statistical properties of any method to handle missing values. This concerns the bias of Complete Case Analysis and other ad hoc methods as well as the efficiency of sophisticated methods.

Consequence of adding a Missing category

Adding a “missing” category can lead to noticeable bias if the categorical covariate is an **important confounder**.

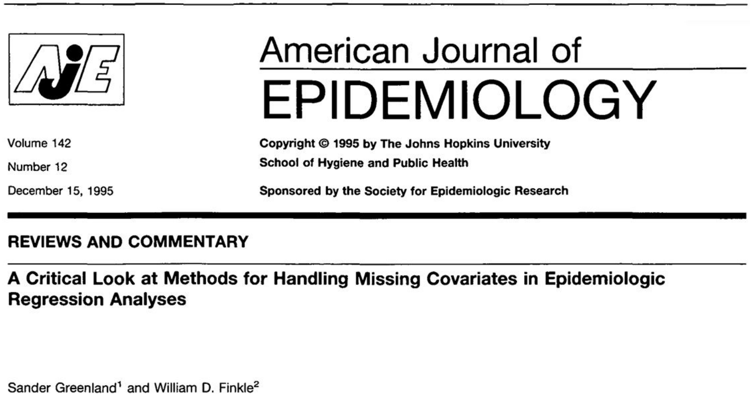


American Journal of Epidemiology
Copyright © 1991 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol 134, No 8
Printed in U S A

Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables

Werner Vach[1] and Maria Blettner[2]



American Journal of EPIDEMIOLOGY

Volume 142
Number 12
December 15, 1995

Copyright © 1995 by The Johns Hopkins University School of Hygiene and Public Health

Sponsored by the Society for Epidemiologic Research

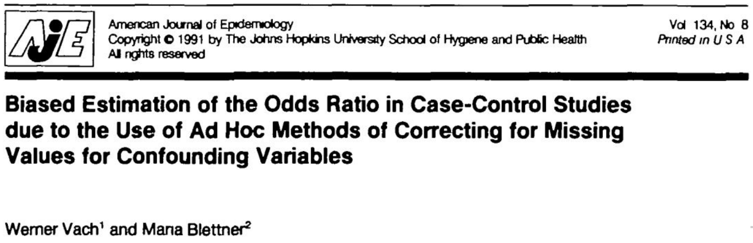
REVIEWS AND COMMENTARY

A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses

Sander Greenland[1] and William D. Finkle[2]

- If a categorical variable e.g., **education level** has missing data, creating a "Missing" category treats the lack of information **as if it's a valid education category** (similar to "High School" or "College"). However, there is no substantive meaning to this "Missing" group in the context of education.

Consequence of adding a Missing category

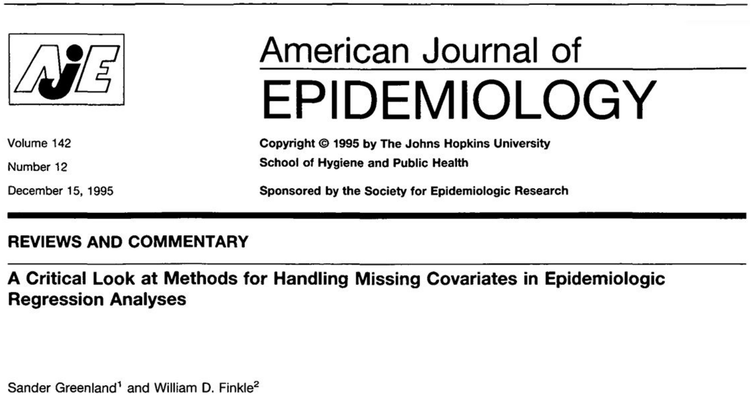


American Journal of Epidemiology
Copyright © 1991 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol 134, No 8
Printed in U S A

Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables

Werner Vach[1] and Maria Blettner[2]



American Journal of
EPIDEMIOLOGY

Volume 142
Number 12
December 15, 1995

Copyright © 1995 by The Johns Hopkins University
School of Hygiene and Public Health

Sponsored by the Society for Epidemiologic Research

REVIEWS AND COMMENTARY

A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses

Sander Greenland[1] and William D. Finkle[2]

- In a study on health outcomes, if **lower-income individuals are more likely to have missing data for their income** (an MNAR scenario, discussed later), **creating a "Missing" category may falsely dilute or mask the relationship between income and health**. As a result, the model might underestimate the effect of income on health outcomes.

When Single imputation (SI) may be preferred

Jakobsen *et al. BMC Medical Research Methodology* (2017) 17:162
DOI 10.1186/s12874-017-0442-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts

Janus Christian Jakobsen[1,2*], Christian Gluud[1], Jørn Wetterslev[1] and Per Winkel[1]

- When only outcome variable is missing and **auxiliary variables** (surrogate / proxy) are available, SI may be better than CC (particularly when variable has substantial amount of missing).
- When missingness is **monotone** (e.g., value only increases), SI can be straightforward (so is MI)
- For clinical trials, SI is often preferred to impute **baseline** covariates.
- For prediction problems, while using **machine learning methods** (e.g., CART) with some more flexibility, but pooling is not straightforward for these approaches (*prediction averaging* is possible as an alternative).

Hot-deck imputation

In single imputation using hot-deck imputation, you are filling the missing data with the response of one person picked at random from a pool of donors **who match on key variables**. You do not take the average of the sample. The **imputed value comes directly from a randomly selected individual from the matched group**, ensuring that the imputed value is a realistic value that exists in the dataset.

Aqua

Dealing with non-normal data

Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

Chapter 3

MVN

- Works with joint model
- Continuous variables only
- **Rubin’s rule was defined under MVN**

MICE

- Works on a variable by variable basis
- One approach: Transform before imputing for non-normal and transform-back after imputation in original scale

Aqua

Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

Dealing with non-normal data

Chapter 3

Transforming may have potential pitfalls:

- **Distortion of relationships between variables** after transformation.
- **Loss of interpretability** of results on the original scale.
- **Inability of some transformations to handle negative or zero values**.
- Back-transformation may **underestimate variance**.

See later regarding alternative methods within MICE.

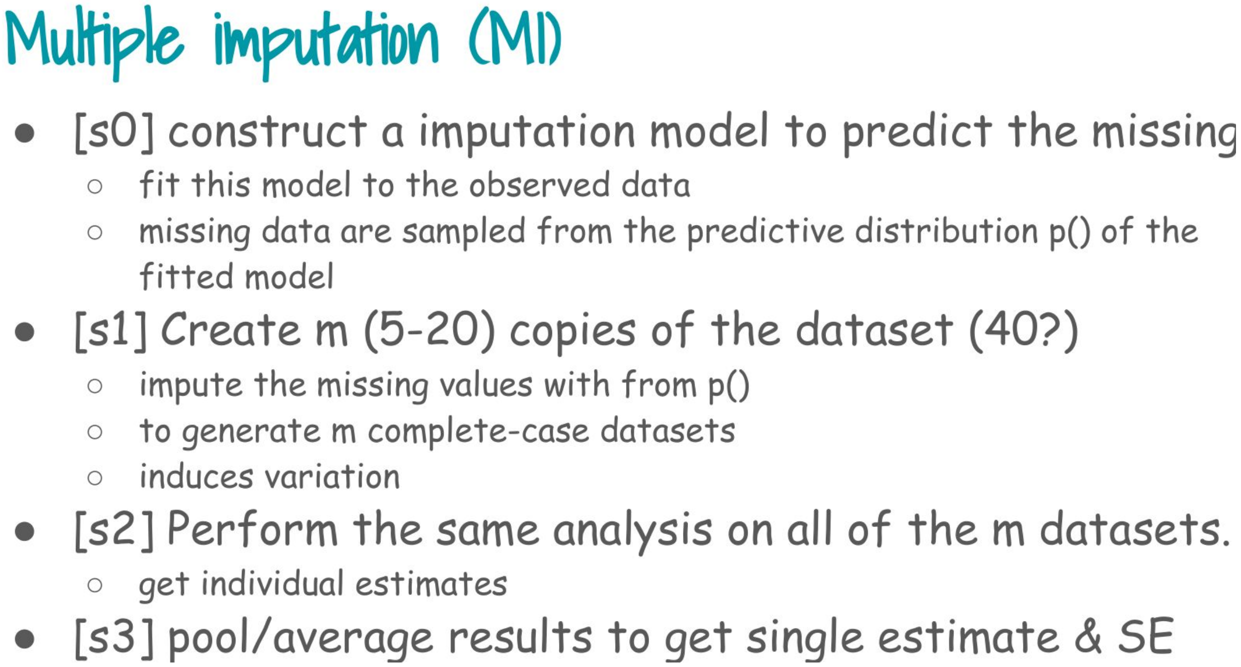
Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

MICE

Steps



Multiple imputation (MI)

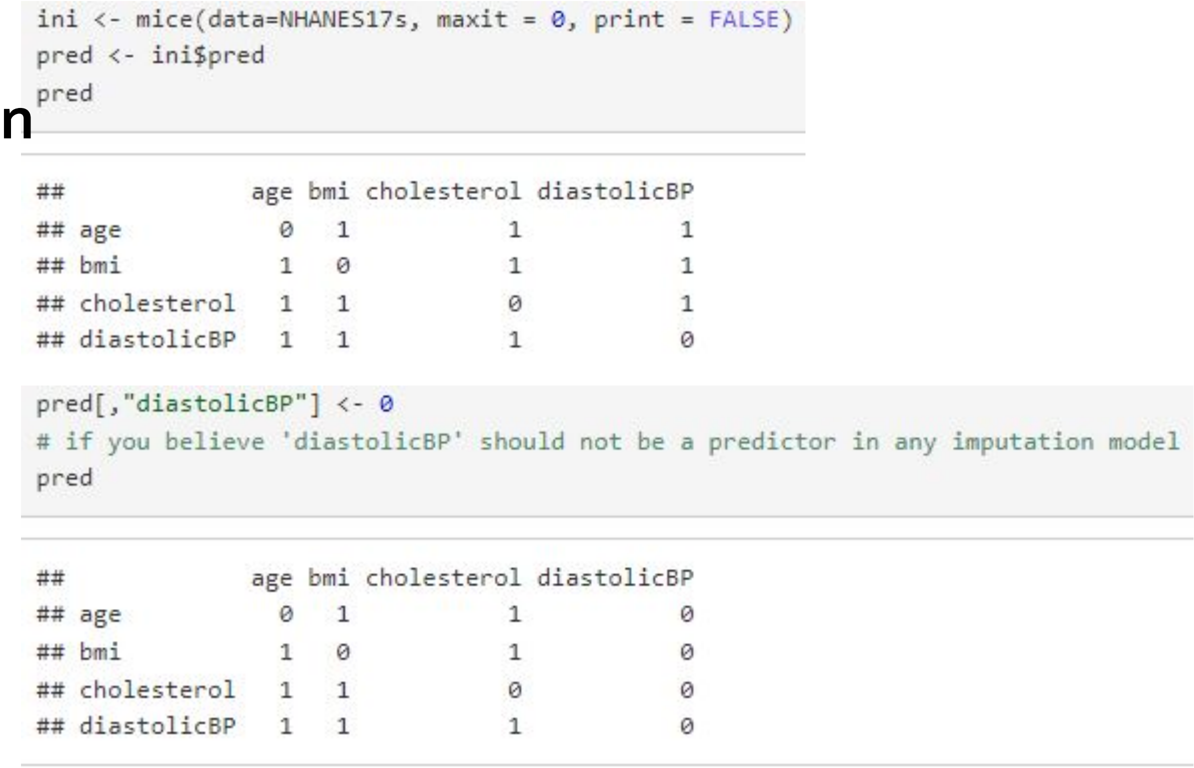
- [s0] construct a imputation model to predict the missing
  - fit this model to the observed data
  - missing data are sampled from the predictive distribution p() of the fitted model
- [s1] Create m (5-20) copies of the dataset (40?)
  - impute the missing values with from p()
  - to generate m complete-case datasets
  - induces variation
- [s2] Perform the same analysis on all of the m datasets.
  - get individual estimates
- [s3] pool/average results to get single estimate & SE

[s0] construct a imputation model to predict the missing

MICE

Implementation

Step 0 (placeholder *[maxit = 0]* imputation model(s))



```
ini <- mice(data=NHANES17s, maxit = 0, print = FALSE)
pred <- ini$pred
pred
```

```
##             age bmi cholesterol diastolicBP
## age           0   1           1           1
## bmi           1   0           1           1
## cholesterol   1   1           0           1
## diastolicBP   1   1           1           0
```

```
pred[,"diastolicBP"] <- 0
# if you believe 'diastolicBP' should not be a predictor in any imputation model
pred
```

```
##             age bmi cholesterol diastolicBP
## age           0   1           1           0
## bmi           1   0           1           0
## cholesterol   1   1           0           0
## diastolicBP   1   1           1           0
```

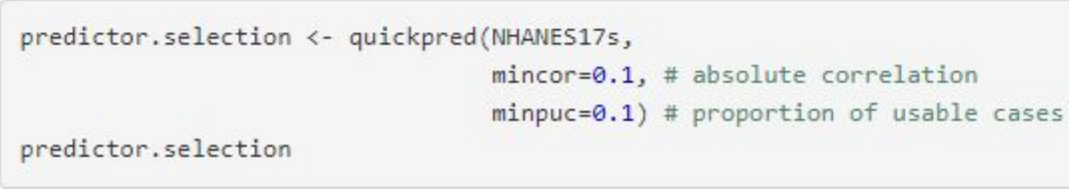
[s0] construct a imputation model to predict the missing

Good to check as an exploratory test; but may be harder to justify if deleting an important known predictor of the imputation target

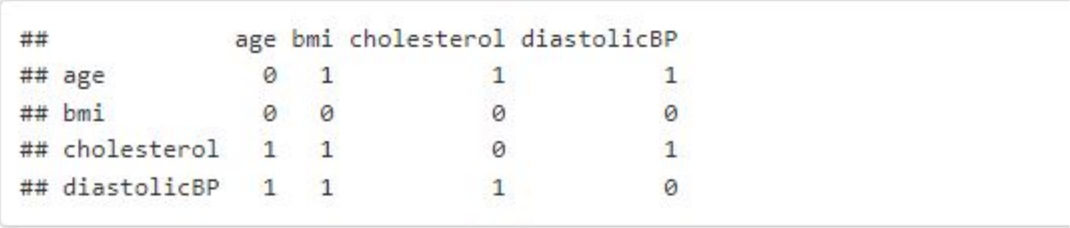
MICE

Implementation

Step 0 (*update* imputation models based on empirical data)



```
predictor.selection <- quickpred(NHANES17s,
                                 mincor=0.1, # absolute correlation
                                 minpuc=0.1) # proportion of usable cases
predictor.selection
```



```
##             age bmi cholesterol diastolicBP
## age           0   1           1           1
## bmi           0   0           0           0
## cholesterol   1   1           0           1
## diastolicBP   1   1           1           0
```

To influence the choice of number of predictors, we can choose different values of

- **mincor** (eliminates predictors whose correlation with imputation model target/outcome is below 0.1) and
- **minpuc** (eliminates predictors whose proportion of usable cases are below 0.1)

(tuning parameters)

[s0] construct a imputation model to predict the missing

MICE

Implementation

Step 0 (*update* imputation models based on empirical data)

Consideration for choosing variables for the imputation model

- ***Imputation model*** should include *all variables and interactions* that will be used in the ***analysis model***
- *outcome variable* of the ***analysis model***
- ***Auxiliary variables*** (those that are not in analysis model; inclusion improves efficiency)
  - variables related to the ***missingness / nonresponse***
  - variables that are **correlated / proxy / surrogate** for the missing variable (**mincor**)
  - **survey feature** variables while using complex survey data
  - Use **component variables** if imputing derived variable (BMI)
- Remove variables with **too many missing** (**minpuc**)

MICE

Potential overfitting / collinearity issues in imputation model building

Implementation

Step 0

- The implementation in mice can detect multicollinearity.
- As a general solution, the algorithm removes one or more predictors from the model.
- You can turn this option off by using the following *mice(..., remove.collinear=FALSE)*

Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

[s0] construct a imputation model to predict the missing

MICE

Implementation

Step 0 (*update* imputation models based on empirical data)

But choosing variable is only one piece of the puzzle of *model building*.

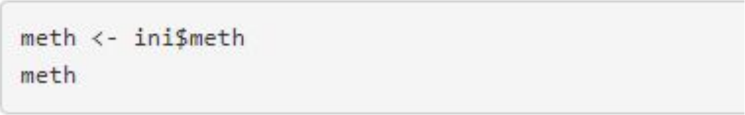
- Is interaction helpful?
- Polynomials?
  - See *mice.impute.quadratic*
- Other transformations? (non-normal?)

[s0] construct a imputation model to predict the missing

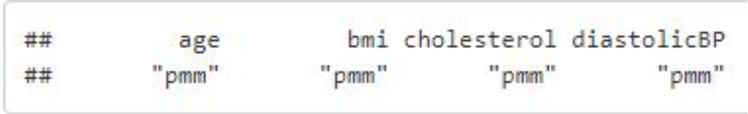
MICE

Implementation

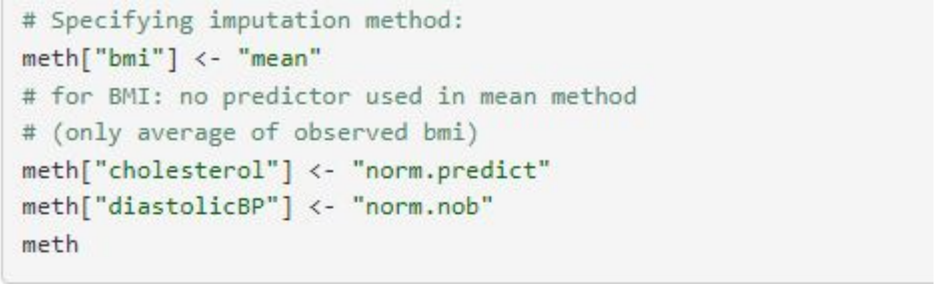
Step 0 (imputation method)



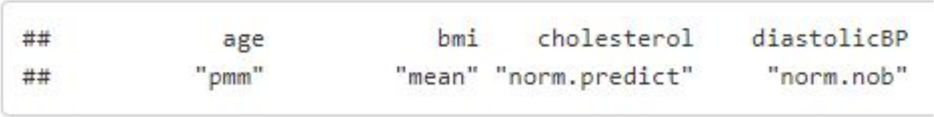
```
meth <- ini$meth
meth
```



```
##         age         bmi cholesterol diastolicBP
##       "pmm"       "pmm"       "pmm"       "pmm"
```



```
# Specifying imputation method:
meth["bmi"] <- "mean"
# for BMI: no predictor used in mean method
# (only average of observed bmi)
meth["cholesterol"] <- "norm.predict"
meth["diastolicBP"] <- "norm.nob"
meth
```



```
##            age            bmi    cholesterol    diastolicBP
##          "pmm"         "mean" "norm.predict"     "norm.nob"
```

MICE methods

Under **MICE, PMM** (method = pmm) is a general / robust strategy within MICE for **non-normal variables**. Since PMM only draws from the observed values, it retains the original data distribution, even if it's skewed or non-normal. It avoids imputing values that don't exist in the data (e.g., extreme or implausible values) and maintains the underlying data characteristics, including skewness or other non-normal features.

Other methods include logistic regression (logreg) or discriminant analysis (lda/qda) for **binary** variables, multinomial logistic regression (polyreg) for **categorical** variables, Poisson regression (poisson) methods for **count** data, with no assumption of normality.

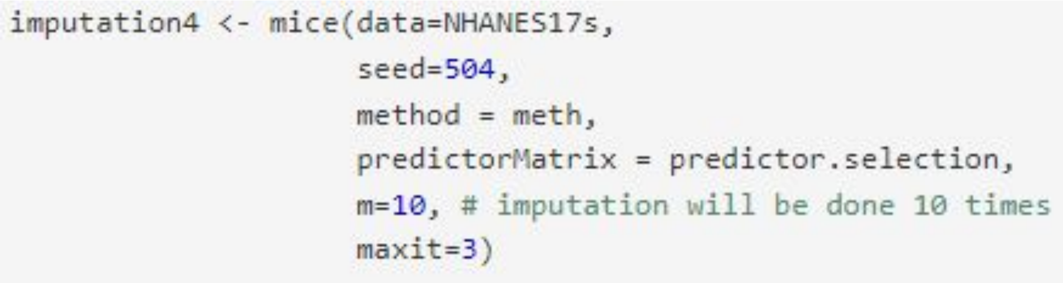
[s1] Create m (5-20) copies of the dataset

MICE

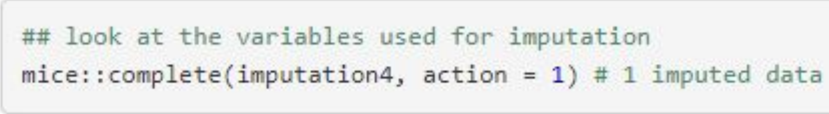
Implementation

Step 1

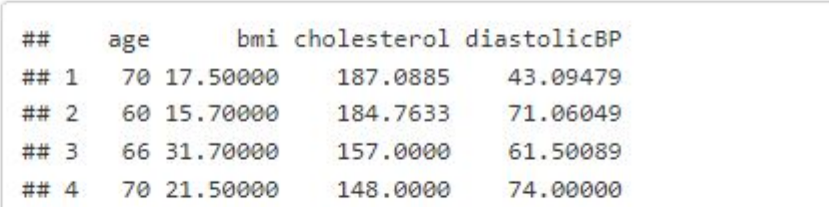
(perform input m times and with a set number of iterations for each imputation)



```
imputation4 <- mice(data=NHANES17s,
                    seed=504,
                    method = meth,
                    predictorMatrix = predictor.selection,
                    m=10, # imputation will be done 10 times
                    maxit=3)
```



```
## look at the variables used for imputation
mice::complete(imputation4, action = 1) # 1 imputed data
```

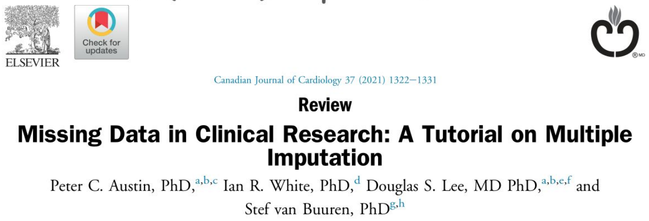


```
##     age      bmi cholesterol diastolicBP
## 1    70 17.50000    187.0885    43.09479
## 2    60 15.70000    184.7633    71.06049
## 3    66 31.70000    157.0000    61.50089
## 4    70 21.50000    148.0000    74.00000
```

[s1] Create m (5-20) copies of the dataset

Van Dyke brown

MICE


Canadian Journal of Cardiology 37 (2021) 1322–1331

Review

Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Peter C. Austin, PhD,[a,b,c] Ian R. White, PhD,[d] Douglas S. Lee, MD PhD,[a,b,e,f] and Stef van Buuren, PhD[g,h]

Implementation

Step 1

(perform input m times and with a set number of iterations for each imputation)

Choosing m

1. 3-5 (Rubin 1987)
2. 5-10 (Schafer SMMR 1999)
3. m should be at least as large as the % of subjects with any missing observations (White Royston Wood, Stat Med 2011)
4. 20-100 (Austin et al CJC 2021)

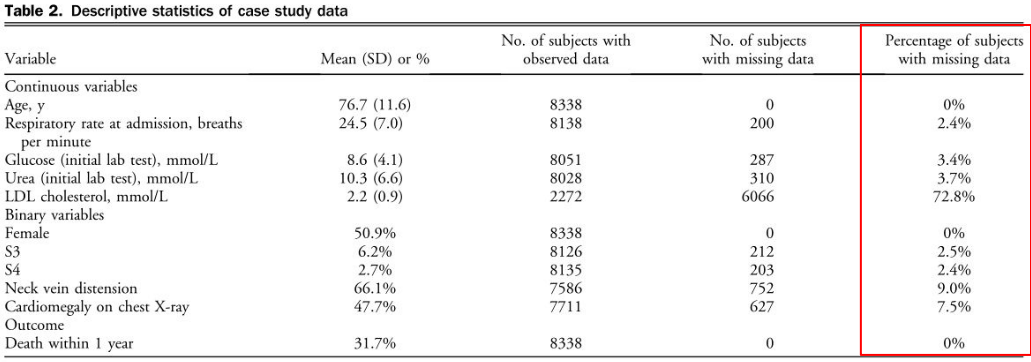


Table 2. Descriptive statistics of case study data

Variable	Mean (SD) or %	No. of subjects with observed data	No. of subjects with missing data	Percentage of subjects with missing data
Continuous variables				
Age, y	76.7 (11.6)	8338	0	0%
Respiratory rate at admission, breaths per minute	24.5 (7.0)	8138	200	2.4%
Glucose (initial lab test), mmol/L	8.6 (4.1)	8051	287	3.4%
Urea (initial lab test), mmol/L	10.3 (6.6)	8028	310	3.7%
LDL cholesterol, mmol/L	2.2 (0.9)	2272	6066	72.8%
Binary variables				
Female	50.9%	8338	0	0%
S3	6.2%	8126	212	2.5%
S4	2.7%	8135	203	2.4%
Neck vein distension	66.1%	7586	752	9.0%
Cardiomegaly on chest X-ray	47.7%	7711	627	7.5%
Outcome				
Death within 1 year	31.7%	8338	0	0%

[s1] Create m (5-20) copies of the dataset

ELSEVIER
Check for updates

Canadian Journal of Cardiology 37 (2021) 1322−1331

Review

Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Peter C. Austin, PhD,[a,b,c] Ian R. White, PhD,[d] Douglas S. Lee, MD PhD,[a,b,e,f] and Stef van Buuren, PhD[g,h]

Amaranth purple

MICE

Implementation

Step 1

(perform input m times and with a set number of iterations for each imputation)

Choosing number of iterations

What does it do?

- In MICE, imputation is done iteratively for each variable with missing values.
- Initially, a **crude imputation (e.g., mean or mode)** is used to fill in missing values for each variable.
- Then, **each variable with missing data is imputed in sequence by using a regression model** based on the other variables in the dataset. This process continues across all variables with missing data.
- After one round of imputation for all variables with missing values, the next iteration (cycle) begins. In e**ach new iteration, the values imputed in previous steps are updated**.
- The **maxit** parameter controls how many of these iterations (cycles) are carried out. Each iteration updates the imputed values as more accurate predictions are made based on the progressively imputed data from earlier steps.

MICE

After a certain number of iterations, the **imputed values typically stabilize**. This means that additional iterations no longer cause substantial changes in the imputed values. This is known as **convergence**.

[s1] Create m (5-20) copies of the dataset

Canadian Journal of Cardiology 37 (2021) 1322–1331

Review

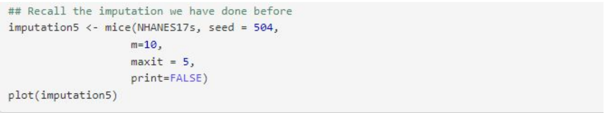
Missing Data in Clinical Research: A Tutorial on Multiple Imputation

Peter C. Austin, PhD,[a,b,c] Ian R. White, PhD,[d] Douglas S. Lee, MD PhD,[a,b,e,f] and Stef van Buuren, PhD[g,h]

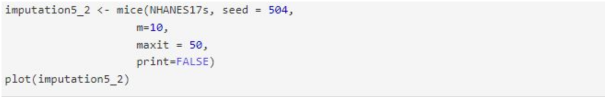
Implementation

Choosing number of iterations

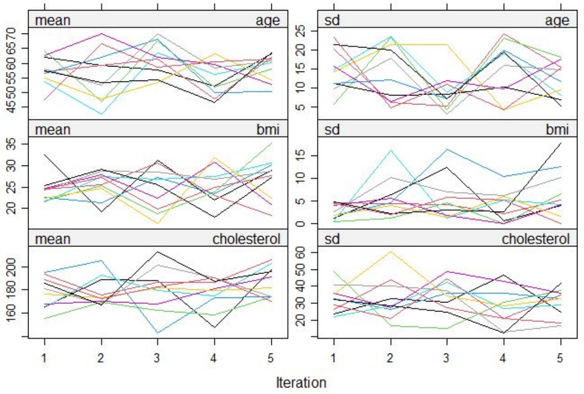
Step 1 (perform input m times and with a set number of iterations for each imputation)

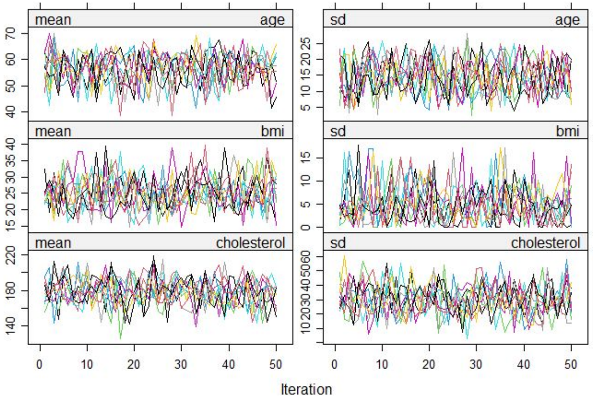


```
## Recall the imputation we have done before
imputation5 <- mice(NHANES17s, seed = 504,
                    m=10,
                    maxit = 5,
                    print=FALSE)
plot(imputation5)
```



```
imputation5_2 <- mice(NHANES17s, seed = 504,
                      m=10,
                      maxit = 50,
                      print=FALSE)
plot(imputation5_2)
```





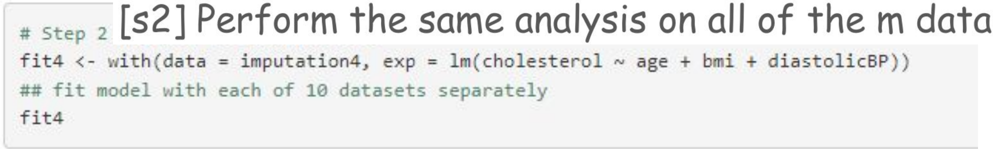
Healthy convergence

MICE

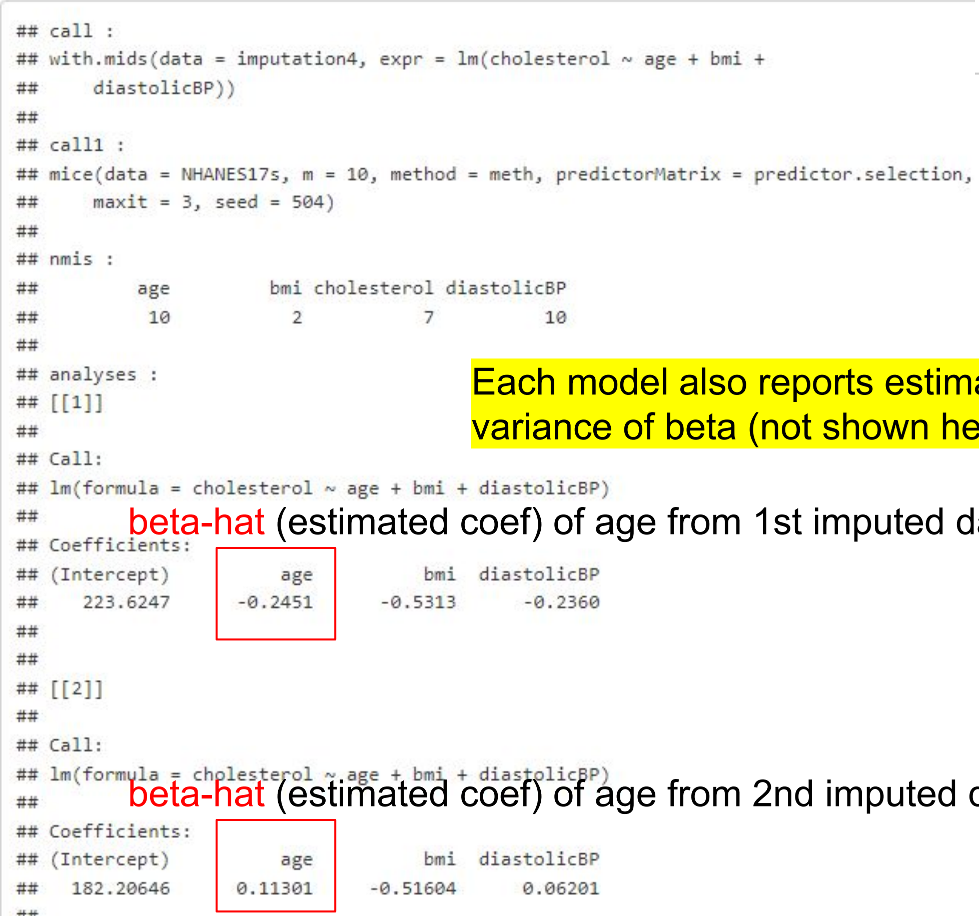
Implementation

Step 2 (analyze m imputed Datasets: results in m estimates)

[s2] Perform the same analysis on all of the m datasets.



```
# Step 2
fit4 <- with(data = imputation4, exp = lm(cholesterol ~ age + bmi + diastolicBP))
## fit model with each of 10 datasets separately
fit4
```



```
## call :
## with.mids(data = imputation4, expr = lm(cholesterol ~ age + bmi +
##     diastolicBP))
##
## call1 :
## mice(data = NHANES17s, m = 10, method = meth, predictorMatrix = predictor.selection,
##     maxit = 3, seed = 504)
##
## nmis :
##         age         bmi cholesterol diastolicBP
##          10           2           7          10
##
## analyses :
## [[1]]
##
## Call:
## lm(formula = cholesterol ~ age + bmi + diastolicBP)
##
## Coefficients:
## (Intercept)          age          bmi  diastolicBP
##    223.6247      -0.2451      -0.5313      -0.2360
##
##
## [[2]]
##
## Call:
## lm(formula = cholesterol ~ age + bmi + diastolicBP)
##
## Coefficients:
## (Intercept)          age          bmi  diastolicBP
##   182.20646      0.11301     -0.51604      0.06201
```

Each model also reports estimated variance of beta (not shown here)

beta-hat (estimated coef) of age from 1st imputed data

beta-hat (estimated coef) of age from 2nd imputed data

[s3] pool/average results to get single estimate & SE

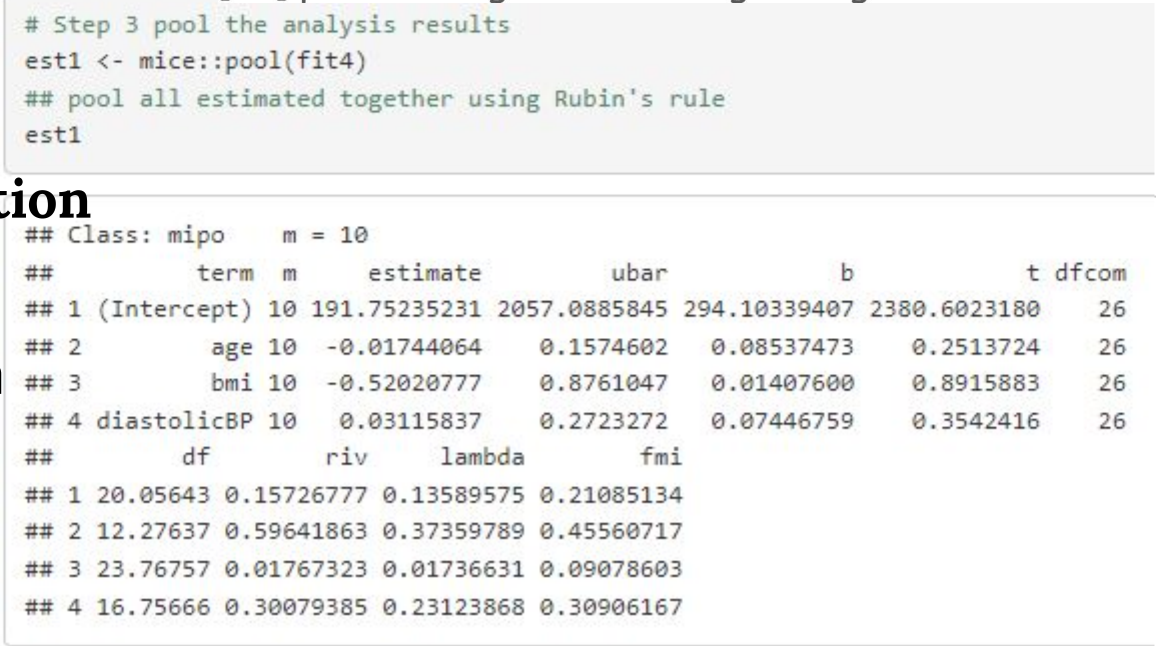
Jazzberry jam

MICE SE Calculation

Implementation

Step 3

(pool)



```
# Step 3 pool the analysis results
est1 <- mice::pool(fit4)
## pool all estimated together using Rubin's rule
est1
```

```
## Class: mipo    m = 10
##          term  m     estimate         ubar           b            t dfcom
## 1 (Intercept) 10 191.75235231 2057.0885845 294.10339407 2380.6023180    26
## 2         age 10  -0.01744064    0.1574602   0.08537473    0.2513724    26
## 3         bmi 10  -0.52020777    0.8761047   0.01407600    0.8915883    26
## 4 diastolicBP 10   0.03115837    0.2723272   0.07446759    0.3542416    26
##          df        riv     lambda        fmi
## 1 20.05643 0.15726777 0.13589575 0.21085134
## 2 12.27637 0.59641863 0.37359789 0.45560717
## 3 23.76757 0.01767323 0.01736631 0.09078603
## 4 16.75666 0.30079385 0.23123868 0.30906167
```

estimate = pooled estimate = sum of (m “beta-hat” estimates) / m (mean of m estimated statistics)

ubar = sum of (m variance[beta] estimates) / m = within-imputation variance (mean of estimated variances)

b = variance of (m “beta-hat” estimates) = between-imputation variance (degree to which estimated statistic / “beta-hat” varies across m imputed datasets). *This b is not available for single imputation when m = 1*.

t = ubar + b + b/m = total variance according to Rubin’s rules (within-imputation & between imputation variation)

dfcom = df for complete

df = Barnard-Rubin correction

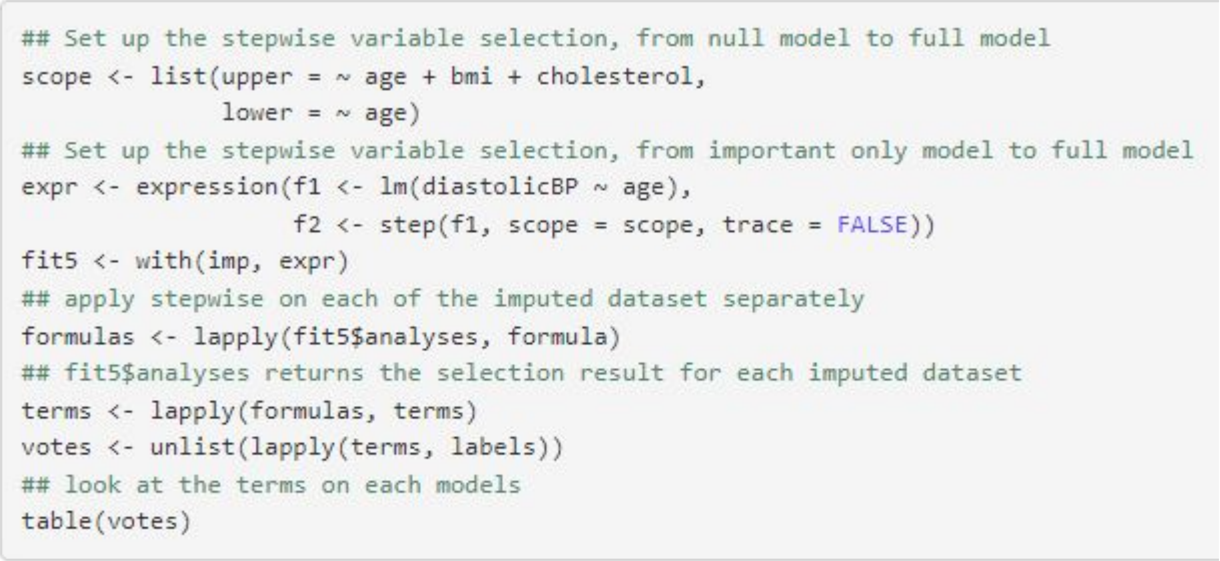
riv = relative increase in variance

lambda = proportion of variance to due nonresponse

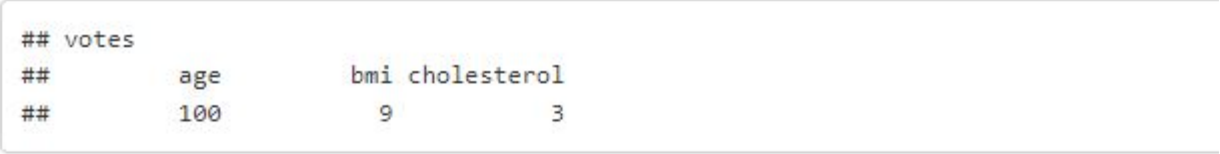
fmi = fraction of missing information per parameter

Variable selection

Majority



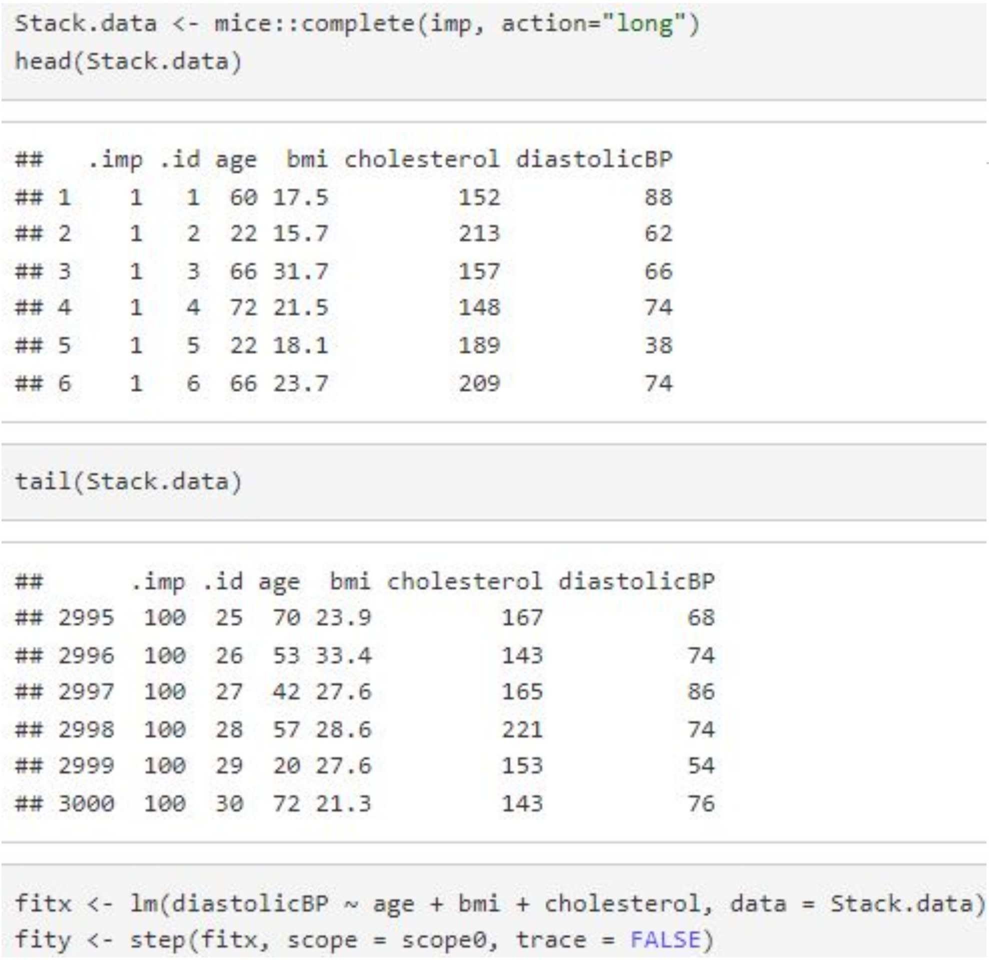
```
## Set up the stepwise variable selection, from null model to full model
scope <- list(upper = ~ age + bmi + cholesterol,
              lower = ~ age)
## Set up the stepwise variable selection, from important only model to full model
expr <- expression(f1 <- lm(diastolicBP ~ age),
                   f2 <- step(f1, scope = scope, trace = FALSE))
fit5 <- with(imp, expr)
## apply stepwise on each of the imputed dataset separately
formulas <- lapply(fit5$analyses, formula)
## fit5$analyses returns the selection result for each imputed dataset
terms <- lapply(formulas, terms)
votes <- unlist(lapply(terms, labels))
## look at the terms on each models
table(votes)
```



```
## votes
##         age         bmi cholesterol
##         100           9           3
```

Variable selection

Stack



```
Stack.data <- mice::complete(imp, action="long")
head(Stack.data)
```

```
##   .imp .id age  bmi cholesterol diastolicBP
## 1    1   1  60 17.5         152          88
## 2    1   2  22 15.7         213          62
## 3    1   3  66 31.7         157          66
## 4    1   4  72 21.5         148          74
## 5    1   5  22 18.1         189          38
## 6    1   6  66 23.7         209          74
```

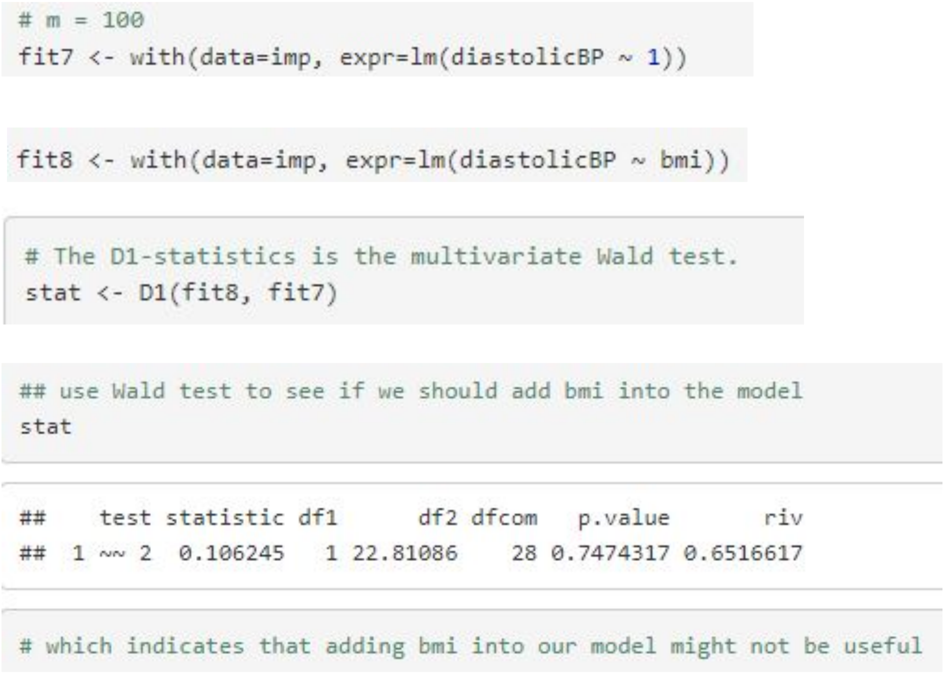
```
tail(Stack.data)
```

```
##      .imp .id age  bmi cholesterol diastolicBP
## 2995  100  25  70 23.9         167          68
## 2996  100  26  53 33.4         143          74
## 2997  100  27  42 27.6         165          86
## 2998  100  28  57 28.6         221          74
## 2999  100  29  20 27.6         153          54
## 3000  100  30  72 21.3         143          76
```

```
fitx <- lm(diastolicBP ~ age + bmi + cholesterol, data = Stack.data)
fity <- step(fitx, scope = scope0, trace = FALSE)
```

Variable selection

Wald



```
# m = 100
fit7 <- with(data=imp, expr=lm(diastolicBP ~ 1))
```

```
fit8 <- with(data=imp, expr=lm(diastolicBP ~ bmi))
```

```
# The D1-statistics is the multivariate Wald test.
stat <- D1(fit8, fit7)
```

```
## use Wald test to see if we should add bmi into the model
stat
```

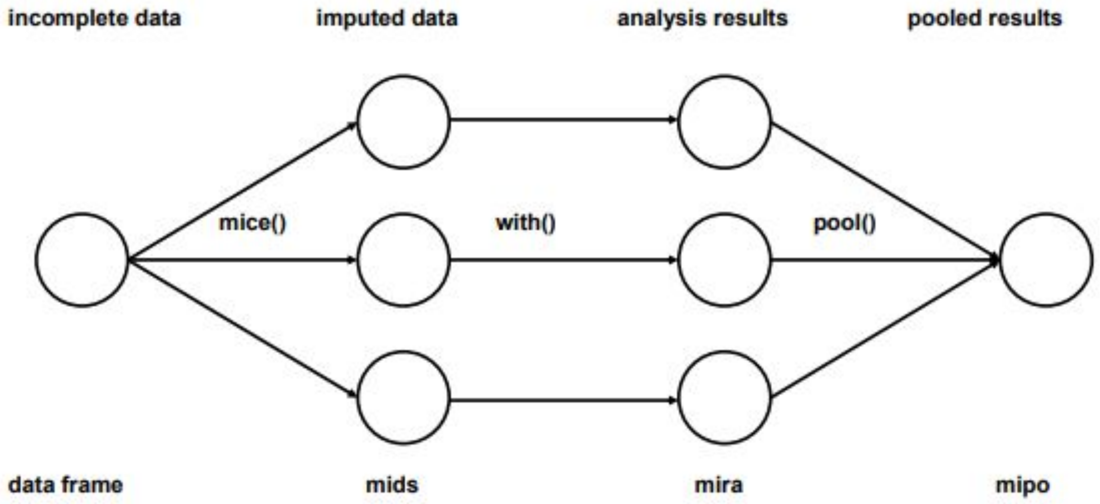
```
##    test statistic df1      df2 dfcom   p.value       riv
##  1 ~~ 2  0.106245   1 22.81086    28 0.7474317 0.6516617
```

```
# which indicates that adding bmi into our model might not be useful
```

Pooling vs. Variable selection

Difference?

- Same model vs. different models in majority rule
- pool



https://ehsanx.github.io/EpiMethods/missingdata3.html

Sheen Green, Amaranth purple

MID for Outcome

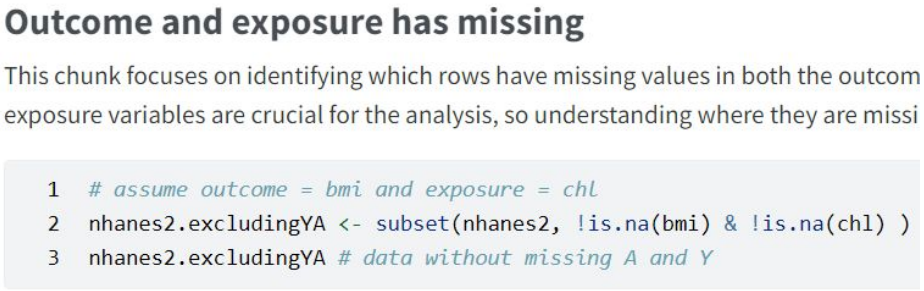
“Multiple imputation **followed by deletion of imputed outcomes** is known as MID. This is very popular, especially when you have high percentage missing values in the outcome variable (e.g., 20%–50%). For **low missing % in outcome, the advantage can be minimal.**”

https://ehsanx.github.io/EpiMethods/missingdata3.html

Sheen Green

MID for Exposure?

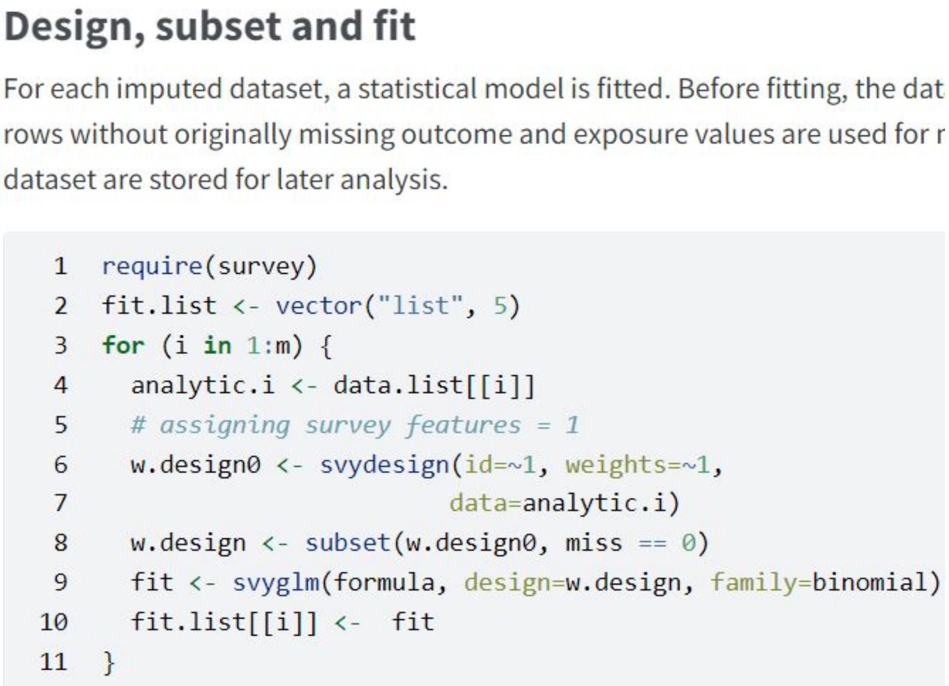
Same idea. See lab above for an example.



Outcome and exposure has missing

This chunk focuses on identifying which rows have missing values in both the outcom
exposure variables are crucial for the analysis, so understanding where they are missi

```r
# assume outcome = bmi and exposure = chl
nhanes2.excludingYA <- subset(nhanes2, !is.na(bmi) & !is.na(chl) )
nhanes2.excludingYA # data without missing A and Y
```



Design, subset and fit

For each imputed dataset, a statistical model is fitted. Before fitting, the dat
rows without originally missing outcome and exposure values are used for r
dataset are stored for later analysis.

```r
require(survey)
fit.list <- vector("list", 5)
for (i in 1:m) {
  analytic.i <- data.list[[i]]
  # assigning survey features = 1
  w.design0 <- svydesign(id=~1, weights=~1,
                        data=analytic.i)
  w.design <- subset(w.design0, miss == 0)
  fit <- svyglm(formula, design=w.design, family=binomial)
  fit.list[[i]] <-  fit
}
```

Amaranth purple

MID as sensitivity

When in doubt (or % in between), you can always assess the robustness of your results. You may consider performing a sensitivity analysis where

- you **impute the Outcome/Exposure** and
- compare results with those where the **Outcome/Exposure was left unimputed**.

This can help gauge the impact of any imputation bias.

MNAR

MNAR means that the probability of data being missing is related to the unobserved (missing) values themselves.

MNAR example

If **sicker patients are more likely to drop out** of a study, their missingness is related to their **health condition**. In this case, the reason a patient drops out (and thus has missing data) is because of their health status. Importantly, **this health status (e.g., their worsening condition or more severe symptoms) is unobserved for those who drop out**. If you tried to explain the dropout using only the observed data (e.g., the **baseline health condition or other demographic characteristics), you might not fully capture the reason for dropout, because it's specifically related to the worsening health condition**, which you don't have data on for those who dropped out.

Why MNAR produces bias?

- Standard statistical methods (e.g., **complete case** analysis or **MAR**-based imputation techniques) assume that missingness is either **random** or **can be predicted by other observed variables**. With MNAR data, this assumption doesn't hold, leading to biased parameter estimates.
- Since the **missingness mechanism depends on the unobserved values**, it's impossible to directly observe the cause of missingness, making it hard to model or adjust for.

Atomic Tangerine

MNAR & subsequent sensitivity analysis

You can impute missing values under different assumptions (e.g., **assume different values for those with missing data: *best health and worst health for the unobserved health condition***) and compare how sensitive your results are to these assumptions.

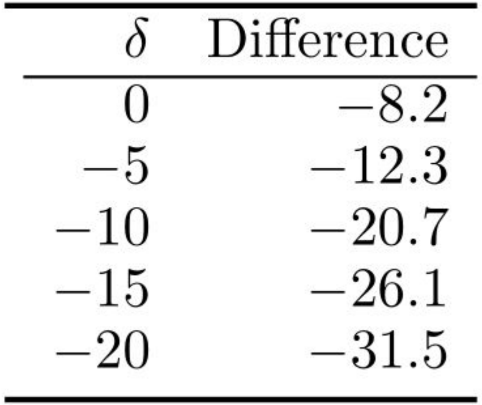
Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

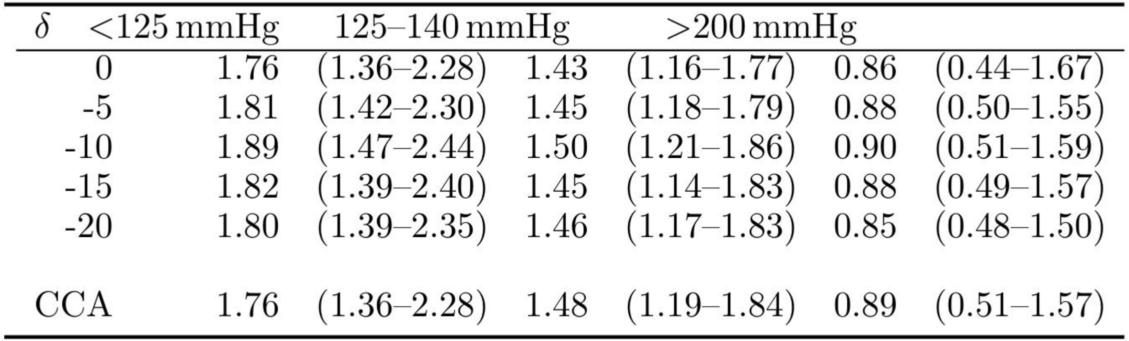
Chapter 9

Delta-adjustment or adding offset in the imputed values



δ	Difference
0	−8.2
−5	−12.3
−10	−20.7
−15	−26.1
−20	−31.5

“differences in means between the imputed and observed
data as a function of delta”



δ	<125 mmHg		125–140 mmHg		>200 mmHg	
0	1.76	(1.36–2.28)	1.43	(1.16–1.77)	0.86	(0.44–1.67)
-5	1.81	(1.42–2.30)	1.45	(1.18–1.79)	0.88	(0.50–1.55)
-10	1.89	(1.47–2.44)	1.50	(1.21–1.86)	0.90	(0.51–1.59)
-15	1.82	(1.39–2.40)	1.45	(1.14–1.83)	0.88	(0.49–1.57)
-20	1.80	(1.39–2.35)	1.46	(1.17–1.83)	0.85	(0.48–1.50)
CCA	1.76	(1.36–2.28)	1.48	(1.19–1.84)	0.89	(0.51–1.57)

“HR estimates under the different scenarios for 3 systolic BP groups”

Thanks!

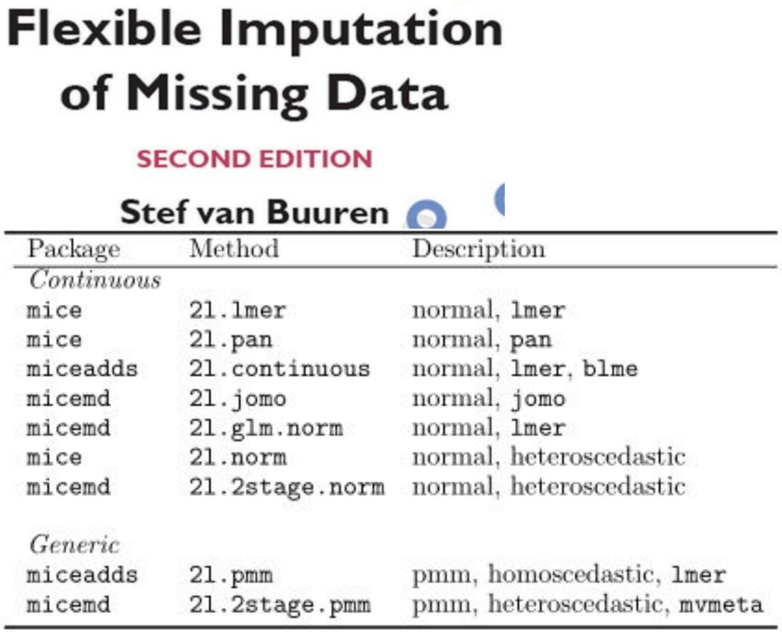
ehsan.karim@ubc.ca

www.ehsank.com

Briefly mentioned in the lecture, but mostly beyond the scope of current course

Multilevel modelling

Many 2l methods developed

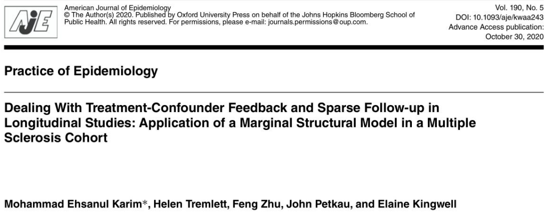


Flexible Imputation of Missing Data

SECOND EDITION

Stef van Buuren

Package	Method	Description
Continuous		
mice	2l.lmer	normal, lmer
mice	2l.pan	normal, pan
miceadds	2l.continuous	normal, lmer, blme
micemd	2l.jomo	normal, jomo
micemd	2l.glm.norm	normal, lmer
mice	2l.norm	normal, heteroscedastic
micemd	2l.2stage.norm	normal, heteroscedastic
Generic		
miceadds	2l.pmm	pmm, homoscedastic, lmer
micemd	2l.2stage.pmm	pmm, heteroscedastic, mvmeta



American Journal of Epidemiology
© The Author(s) 2020. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 190, No. 5
DOI: 10.1093/aje/kwaa243
Advance Access publication:
October 30, 2020

Practice of Epidemiology

Dealing With Treatment-Confounder Feedback and Sparse Follow-up in Longitudinal Studies: Application of a Marginal Structural Model in a Multiple Sclerosis Cohort

Mohammad Ehsanul Karim*, Helen Tremlett, Feng Zhu, John Petkau, and Elaine Kingwell

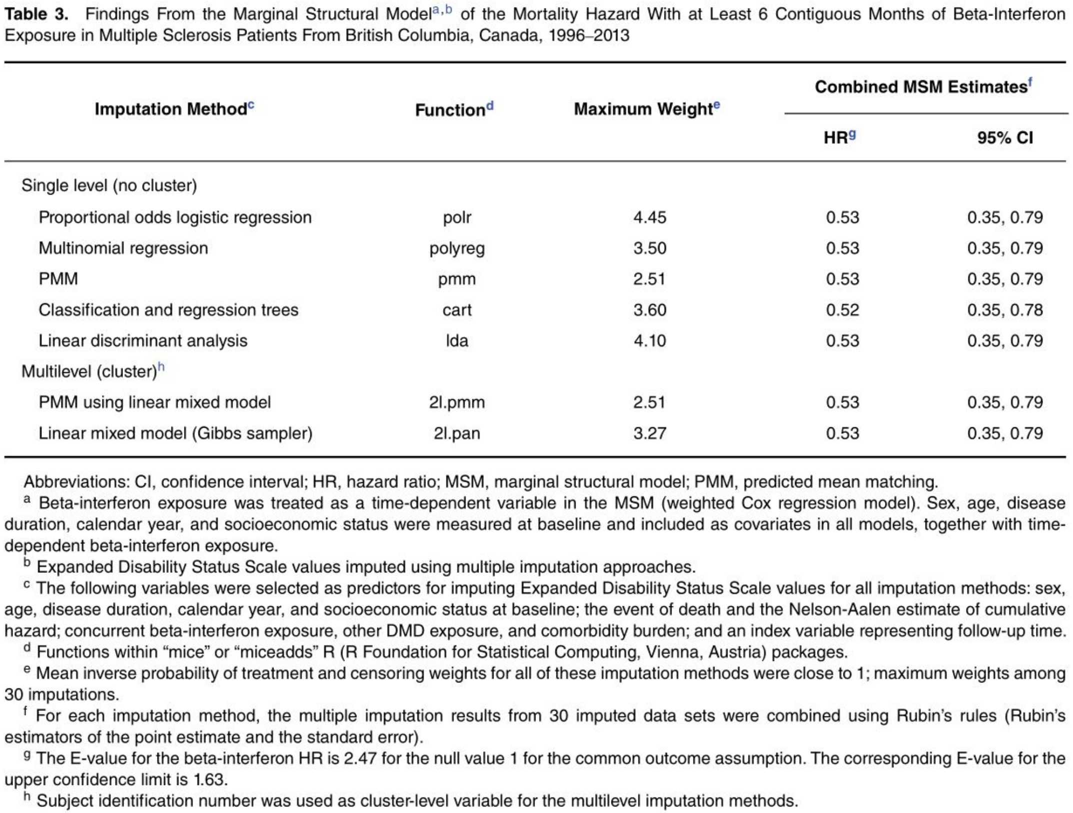


Table 3. Findings From the Marginal Structural Model[a,b] of the Mortality Hazard With at Least 6 Contiguous Months of Beta-Interferon Exposure in Multiple Sclerosis Patients From British Columbia, Canada, 1996–2013

Imputation Method[c]	Function[d]	Maximum Weight[e]	Combined MSM Estimates[f]	
			HR[g]	95% CI
Single level (no cluster)				
Proportional odds logistic regression	polr	4.45	0.53	0.35, 0.79
Multinomial regression	polyreg	3.50	0.53	0.35, 0.79
PMM	pmm	2.51	0.53	0.35, 0.79
Classification and regression trees	cart	3.60	0.52	0.35, 0.78
Linear discriminant analysis	lda	4.10	0.53	0.35, 0.79
Multilevel (cluster)[h]				
PMM using linear mixed model	2l.pmm	2.51	0.53	0.35, 0.79
Linear mixed model (Gibbs sampler)	2l.pan	3.27	0.53	0.35, 0.79

Abbreviations: CI, confidence interval; HR, hazard ratio; MSM, marginal structural model; PMM, predicted mean matching.

[a] Beta-interferon exposure was treated as a time-dependent variable in the MSM (weighted Cox regression model). Sex, age, disease duration, calendar year, and socioeconomic status were measured at baseline and included as covariates in all models, together with time-dependent beta-interferon exposure.

[b] Expanded Disability Status Scale values imputed using multiple imputation approaches.

[c] The following variables were selected as predictors for imputing Expanded Disability Status Scale values for all imputation methods: sex, age, disease duration, calendar year, and socioeconomic status at baseline; the event of death and the Nelson-Aalen estimate of cumulative hazard; concurrent beta-interferon exposure, other DMD exposure, and comorbidity burden; and an index variable representing follow-up time.

[d] Functions within "mice" or "miceadds" R (R Foundation for Statistical Computing, Vienna, Austria) packages.

[e] Mean inverse probability of treatment and censoring weights for all of these imputation methods were close to 1; maximum weights among 30 imputations.

[f] For each imputation method, the multiple imputation results from 30 imputed data sets were combined using Rubin's rules (Rubin's estimators of the point estimate and the standard error).

[g] The E-value for the beta-interferon HR is 2.47 for the null value 1 for the common outcome assumption. The corresponding E-value for the upper confidence limit is 1.63.

[h] Subject identification number was used as cluster-level variable for the multilevel imputation methods.