


Adv. Stat. Topics A - Missing data  
Morning session


Anne Helby Petersen

# Program outline

- 08.15-09.00: The Elderly study: An example of a study with missing information
- 09.00-09.10: Break
- 09.10-09.45: Missing information in R & classification of missing information
- 09.45-10.40: Work with data: Describe missing information
- 10.40-11.00: Presentations
- 11.00-12.00: Lunch
- 12.00-12.50: Imputation & Multiple Imputation using Chained Equations (MICE)
- 12.50-14.15: Work with data: Data analysis with missing information
- 14.15-14.45: Presentations
- 14.45-15.00: Further perspectives and more resources

Research Report |  Full Access |

### **Evaluation of adding the Community Reinforcement Approach to Motivational Enhancement Therapy for Adults Aged 60 Years and Older with DSM-5 Alcohol Use Disorder: A Randomised Controlled Trial**

Kjeld Andersen , Silke Behrendt, Randi Bilberg, Michael P. Bogenschütz, Barbara Braun, Gerhard Buehringer, Claus Thorn Ekstrøm, Anna Mejldal, Anne Helby Petersen, Anette Søgaard Nielsen

First published: 27 August 2019 | <https://doi.org/10.1111/add.14795>

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/add.14795.

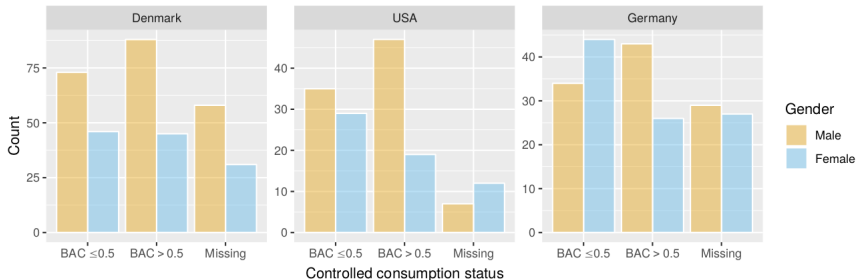
## The Elderly study - overview

- ▶ RCT on 693 patients from USA, DK and DE, suffering from alcohol dependency, all 60+ years old.
- ▶ Purpose: Compare usual treatment (MET) with new treatment that contains an additional element specifically designed for this elderly population (MET+CRA).
- ▶ Primary outcome: Controlled consumption (CC) status after approx. 6 months of treatment (blood alcohol level  $\leq 0.05\%$  at all times during 30 days).
- ▶ Seven baseline covariates included to reduce noise in the models: country (DK, DE, USA), gender (male/female), age (measured as years older than 60), education (no degree/at most undergrad./grad. or post grad.), cohabiting with partner (yes/no), alcohol dependence severity (low/intermediate/severe), number of previous treatments (0/1-2/3+)

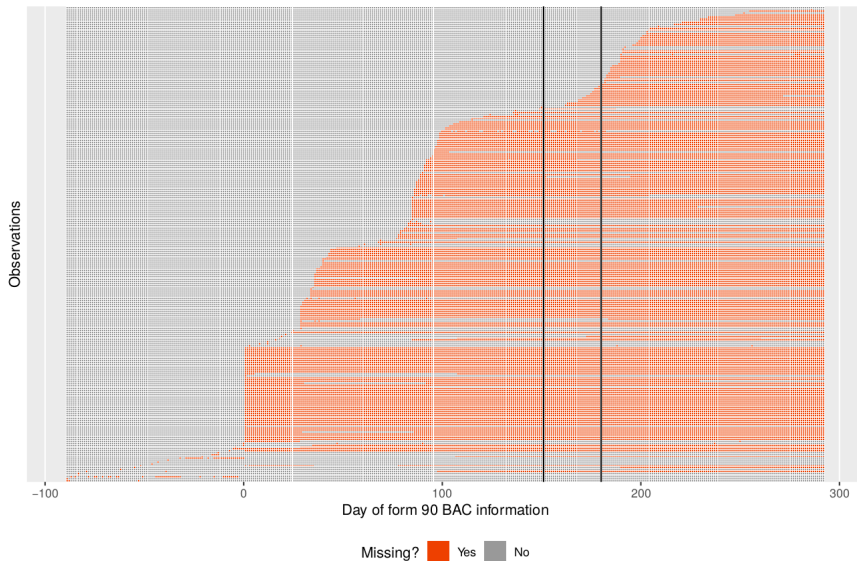
# The Elderly study - missing information

Missing information in two variables:

- ▶ Dependency severity: 3 patients (0.43%) (*excluded*)
- ▶ Controlled consumption status: 164 patients (23.77%)



# The Elderly study - closer look at missingness in CC status



# The Elderly study - non-response analysis

We fitted a logistic regression model with:

**Outcome:** Indicator of whether the patient has missing CC status

**Predictors:** All the remaining variables

We tested significance of each of the predictors:

	Change in df	AIC	LRT	p-value
Full model		740.8		
Treatment	1	739.2	0.3952	0.5296
Gender	1	738.8	0.0032	0.9551
Country	2	748.9	12.0718	0.0024
Age	1	746.6	7.8397	0.0051
Education	2	741.8	4.9461	0.0843
Partner status	1	744.5	5.7452	0.0165
Alcohol dependence severity	2	739.1	2.3205	0.3134
Previous treatment history	2	744.7	7.8811	0.0194

# The Elderly study - statistical methods

- ▶ We estimated the difference between the two treatments using logistic regression.
- ▶ The primary model was fitted on complete cases only.
- ▶ In sensitivity analyses, we compared this model to
  - ▶ A model using multiple imputation
  - ▶ Several best case/worst case scenario models



# The Elderly study - results

Table 3.1: Estimated log odds ratios from the model of controlled consumption status using all full covariate adjustment. The reported estimates are on log odds ratio scale and they are computed relative to the following reference category: Treatment MET; Gender male; Country Denmark; Age 60; Education none; No partner; Low ADS; Previous treatments 0. The mean log odds of having a controlled alcohol consumption in this reference group is represented by the intercept estimate. The reported p-values correspond to two-sided z-tests of the null-hypothesis of a zero parameter value.

	Estimate	Std. error	z statistic	p-value
<b>Intercept</b>	-0.3507	0.3050	-1.1499	0.2502
<b>Treatment: MET+CRA</b>	0.2028	0.1801	1.1260	0.2602
<b>Country: USA</b>	0.0736	0.2327	0.3164	0.7517
<b>Country: Germany</b>	-0.0351	0.2522	-0.1392	0.8893
<b>Gender: Female</b>	-0.5543	0.1906	-2.9085	0.0036
<b>Age</b>	0.0677	0.0211	3.2038	0.0014
<b>Married or cohabiting: Yes</b>	0.2270	0.1877	1.2094	0.2265
<b>Severity: Intermediate</b>	-0.0777	0.2307	-0.3367	0.7363
<b>Severity: Substantial or severe</b>	-0.2767	0.4096	-0.6755	0.4994
<b>Education: At most undergraduate degree</b>	0.0518	0.2286	0.2268	0.8206
<b>Education: Graduate or post-graduate</b>	-0.4463	0.2872	-1.5537	0.1202
<b>Previous treatments: 1-2</b>	0.2655	0.2187	1.2140	0.2247
<b>Previous treatments: 3+</b>	0.2938	0.3087	0.9517	0.3413

# Hypothetical long-term follow-up study

- ▶ Outcome: Diagnosis of liver disease within 10 years, measured in national registers (assume no censoring, no death).
- ▶ Explanatory variable of interest: Controlled consumption status after 6 months of treatment.
- ▶ Other explanatory variables as before: Country, gender, age, education, partner status, alcohol dependency severity, previous treatment history.
- ▶ What you see in the data: 24% of the observations having missing CC information.

## 24% of the patients have missing CC information...

**Scenario 1:** ... due to a fire in the storing facility.

**Scenario 2:** ... because those patients were embarrassed to tell the treatment facility that they had started drinking again.

**Scenario 3:** ... and they are the 24% of the patients with the most severe alcohol dependencies, and they dropped out of the study.

**Scenario 4:** ... and they are the 24% who are female, and they dropped out of the study.

**Scenario 5:** ... because those patients all had last names starting with "A" and their records were lost because someone dropped a cup of coffee on the folder that contained them.

**Scenario 6:** ... because those patients dropped out of the study since they were not drinking and felt safe that they wouldn't start again.

**Scenario 7:** ... and they are the 24% that have red hair. They are missing in CC because a data manager accidentally deleted their information - and the variable containing hair color.

**Scenario 8:** ... and they all had undiagnosed pre-stages to liver disease during the study and dropped out due to illness.

## Discuss the missing information scenarios

Assume that we carry out a statistical analysis (e.g. logistic regression model) using only the patients with no missing information (*complete case analysis*).

For each of the eight scenarios, discuss with your neighbors:

- ▶ Will this affect the estimate of the effect of CC status on liver disease risk?
- ▶ Will this affect the precision (e.g. wideness of confidence intervals) of our effect estimate?
- ▶ Can this problem be solved using statistical methods? And do you have any suggestions for how?

We follow up afterwards.

First rule of missing information handling in R:  
Always represent missing values by NA.

First rule of missing information handling in R:  
Always represent missing values by NA.

Second rule of missing information handling in R:  
**Always represent missing values by NA.**

## A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")

## A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")



## A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")
3. A blank space (" ")

## A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")
3. A blank space (" ")
4. A dash ("-")

## A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")
3. A blank space (" ")
4. A dash ("-")
5. A special numeric value (e.g. Inf (infinity) or NaN (not a number))

# A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")
3. A blank space (" ")
4. A dash ("-")
5. A special numeric value (e.g. Inf (infinity) or NaN (not a number))
6. An unusual numeric value (e.g. 999, -9, 88, ...)

# A list of bad ideas

Do *not* represent missing information by:

1. A dot (".")
2. An empty character string ("")
3. A blank space (" ")
4. A dash ("-")
5. A special numeric value (e.g. Inf (infinity) or NaN (not a number))
6. An unusual numeric value (e.g. 999, -9, 88, ...)
7. Anything else that is not NA.

## A quick check for bad missing data

If someone gave you bad data, you can use the `dataReporter` package in R to look for problems:

```
> library(dataReporter)
> testData$miscodedMissingVar
[1] "."      ""      "nan"   "NaN"   "NAN"   "na"    "NA"
[7] "Na"     "Inf"   "inf"   "-Inf"  "-inf"  "-"      "9"
[15] "9"
```

## A quick check for bad missing data

If someone gave you bad data, you can use the `dataReporter` package in R to look for problems:

```
> library(dataReporter)
> testData$miscodedMissingVar
[1] "."      ""      "nan"   "NaN"   "NAN"   "na"    "NA"
[7] "Na"     "Inf"   "inf"   "-Inf"  "-inf"  "-"     "9"
[15] "9"
```

```
> identifyMissing(testData$miscodedMissingVar)
```

The following suspected missing value codes enter as regular values: `,` `-`, `-inf`, `-Inf`, `.`, `9`, `inf`, `Inf`, `na`, `Na` (4 additional values omitted).

## A quick check for bad missing data

If someone gave you bad data, you can use the `dataReporter` package in R to look for problems:

```
> library(dataReporter)
> testData$miscodedMissingVar
[1] "."      ""      "nan"   "NaN"   "NAN"   "na"    "NA"
[7] "Na"    "Inf"   "inf"   "-Inf"  "-inf"  "-"     "9"
[15] "9"
```

```
> identifyMissing(testData$miscodedMissingVar)
```

The following suspected missing value codes enter as regular values: `,` `-`, `-inf`, `-Inf`, `.`, `9`, `inf`, `Inf`, `na`, `Na` (4 additional values omitted).

Or get a full report with more general checks for problems in the data:

```
> library(dataReporter)
> makeDataReport(testData)
```



## Back to theory: An evil scheme

Imagine we had a fully observed dataset, but wish to induce missing information in one variable. How can we make data go missing?

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

# Missing completely at random (MCAR)

**Evil scheme:** Choose who is missing by random dice roll.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

**Why is this evil?**

# Missing completely at random (MCAR)

**Evil scheme:** Choose who is missing by random dice roll.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

**Why is this evil?** We lose information and hence precision (wider confidence intervals).

# Missing at random (MAR)

**Evil scheme:** Choose who is missing by separate random draws for males and females. Females have missing probability 0.75, males have missing probability 0.05.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	
14	Denmark	Male	0	No

**Why is this evil?**

# Missing at random (MAR)

**Evil scheme:** Choose who is missing by separate random draws for males and females. Females have missing probability 0.75, males have missing probability 0.05.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	
14	Denmark	Male	0	No

**Why is this evil?** Underrepresentation of females may lead to biased estimates.

# Missing not at random (MNAR)

**Evil scheme:** Choose who is missing by looking at CC status itself. Relapsers are more likely to have missing information.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

**Why is this evil?**

# Missing not at random (MNAR)

**Evil scheme:** Choose who is missing by looking at CC status itself. Relapsers are more likely to have missing information.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No



	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

**Why is this evil?** Underrepresentation of relapsers may lead to biased estimates. And the worst part: *Whether an observation is missing depends on the very information we are missing.*

# Three categories of missing information

Missing information can be divided into these three categories (*yes, those are the actual standard names*):

- MCAR Missing completely at random:** Whether an observation is missing does not depend on observed, nor unobserved, variables. (*Dice roll*)
- MAR Missing at random:** Whether an observation is missing does not depend on unobserved variables, but does depend on observed ones. (*Separate dice rolls for groups corresponding to other variables*)
- MNAR Missing not at random:** Whether an observation is missing depends on unobserved variables (and possibly also observed ones). (*Separate dice rolls for groups corresponding to variables that have missing information*)



## Returning to the 8 scenarios

We will now:

- ▶ Classify each of the 8 scenarios as MCAR/MAR/MNAR
- ▶ Discuss: Would it had been possible to detect this by looking at the data alone (i.e. not knowing why the data went missing)?

## 24% of the patients have missing CC information...

**Scenario 1:** ... due to a fire in the storing facility.

**Scenario 2:** ... because those patients were embarrassed to tell the treatment facility that they had started drinking again.

**Scenario 3:** ... and they are the 24% of the patients with the most severe alcohol dependencies, and they dropped out of the study.

**Scenario 4:** ... and they are the 24% who are female, and they dropped out of the study.

**Scenario 5:** ... because those patients all had last names starting with "A" and their records were lost because someone dropped a cup of coffee on the folder that contained them.

**Scenario 6:** ... because those patients dropped out of the study since they were not drinking and felt safe that they wouldn't start again.

**Scenario 7:** ... and they are the 24% that have red hair. They are missing in CC because a data manager accidentally deleted their information - and the variable containing hair color.

**Scenario 8:** ... and they all had undiagnosed pre-stages to liver disease during the study and dropped out due to illness.

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- ▶ Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- ▶ Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

## Strategy 2:

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- ▶ Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

## Strategy 2: NA

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- ▶ Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

## Strategy 2: NA

- ▶ Nothing more can be done using data and statistics alone.

# Distinguishing between MCAR/MAR/MNAR with data & statistics

## Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- ▶ Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

## Strategy 2: NA

- ▶ Nothing more can be done using data and statistics alone.

**Conclusion: The only "test" you can perform is to falsify a MCAR assumption.** Distinguishing between MAR and MNAR *must* be based on discussion, sensitivity analyses and external knowledge (more on that in the afternoon).



## Data exercise: Looking for trouble

- ▶ We will now start looking at data with missing information.
- ▶ We have a dataset consisting of the baseline covariates from the Elderly study and an additional variable, `drinks`, with the mean number of drinks consumed per day in the month before the study started.
- ▶ We wish to model how `drinks` depends on the other baseline covariates.
- ▶ However, an evil person (me) made some of the data go missing.
- ▶ Today's goal is to find out what happened to the data and try to obtain a correct analysis despite the evil scheme.

# Collective Google Slides

- ▶ We will discuss your findings later by help of collectively made Google Slide shows (link in exercises).
- ▶ Along the way: Please add plots/tables/points/drawings/whatever information you want.

## Data exercise: Get started

Go to “Exercise: Explore” on the course website

<https://biostatistics.dk/teaching/advtopicsA/notes.html>

and work through the questions in small groups. We will discuss your findings (using the Google Slide show) around 10:40.