

# Evaluation Randomisiert-Kontrollierter Studien und Experimente mit R

## Missing Data

---

Prof. Dr. David Ebert & Mathias Harrer

Graduiertenseminar TUM-FGZ

Psychology & Digital Mental Health Care, Technische Universität München

## Missing Data-Mechanismen

---

*“Obviously the best way to treat missing data is not to have them.”*

Orchard & Woodbury (1972)

Fehlende Werte sind aber, insbesondere in der medizinischen und psychologischen Forschung, häufig **unvermeidbar!**

## Taxonomie: Gründe für fehlende Werte in klinischen Studien

- **Instrumente:** Response Burden, zu langes Assessment, ...
- **Teilnehmende:** Überforderung, Privatsphäre-Bedenken, Motivationsprobleme, ...
- **Center:** Inadäquate Umsetzung des Studienmethodik, Personalmangel, ...
- **Personal:** Falsche Datenerfassung, Datenverlust, falsche Dateneingabe, ...
- **Studie:** Technische Fehler, Zeitverzögerungen bei Follow-Ups, ...

Palmer et al. (2018)

**Der Umgang mit fehlenden Werten in der (medizinisch-psychologischen) Forschung ist oftmals mangelhaft:** (Akl et al., 2015; Bell et al., 2014; Van Buuren, 2018, Kapitel 1.1.2; Wood et al., 2004)

- Verteilung fehlender Werte nicht transparent berichtet
- Unpassendes missing data handling (z.B. **listwise deletion**)
- Adäquate Imputationsmethoden (z.B. MI, FIML) häufig nicht benutzt, inadäquat angewendet, oder unzureichend berichtet

**R macht es nicht (automatisch) “richtig!”**

```
y <- 1:10
x <- c(1, NA, NA, NA, 3, 5, 8, 10, -1, 10)
summary(lm(y ~ x))
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9403      1.7738   2.785   0.0387 *
## x             0.3172      0.2709   1.171   0.2945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.904 on 5 degrees of freedom
## (3 observations deleted due to missingness)
```

## Was dagegen tun?

- Umsetzung von **Reporting-Standards** z.B. STROBE (Von Elm et al., 2007) & CONSORT (Schulz et al., 2010), s.a. Sterne et al. (2009).
- Adäquates **Missing Data Handling**
  - Basierend auf plausiblen Annahmen, warum fehlende Werte entstanden sind (“Missing Data Mechanism”)
  - ggf. Schätzung (Imputation) fehlender Werte unter Einbezug von deren Unsicherheit

## **MCAR, MAR & MNAR**

---





Donald B. Rubin

**Grundannahme:** Das Fehlen oder Vorhandensein von Daten ist Resultat eines probabilistischen Prozesses.

Diesem Prozess versucht man sich durch ein Modell (**missing data model**) anzunähern.

(Rubin, 1976)



(Harrer et al., 2021)

Nach Rubin (1976) können Missing Data-Mechanismen in 3 Untertypen klassifiziert werden:

## **MCAR**

Missing Completely At Random: rein zufällig fehlende Werte einer Variable.

## **MAR**

Missing At Random: das Fehlen von Werten einer Variable ist abhängig von anderen (observierten) Variablen.

## **MNAR**

Missing Not At Random / "Nonignorable Missing Data": das Fehlen von Werten einer Variable ist (u.A.) abhängig von den Werten der Variable selbst.

→ Für jede Annahme ergeben sich unterschiedliche Auswirkungen bei der Datenauswertung!

**Notation** (Van Buuren, 2018, Kapitel 2.2.3 & 2.2.4)

- $\mathbf{Y}$ :  $n \times p$  Matrix mit teils fehlenden Werten ( $n$  Personen,  $p$  Variablen).
- $\mathbf{X}$ : Matrix mit (vollständig observierten) Kovariaten.
- $\mathcal{D} = (\mathbf{Y}, \mathbf{X})$ : Gesamter Datensatz.
- $\mathbf{R}$ :  $n \times p$  Matrix mit 0 (Datenpunkt fehlt) und 1 (Datenpunkt observiert; “response indicator”).
- $\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}$ : Observierte Daten, fehlende Daten.
- $\psi$ : Parameter des Missing Data-Modells (typischerweise nicht für die wiss. Fragestellung selbst relevant).

→ Missing Data-Modelle treffen Aussagen darüber, **in welcher Beziehung**  $\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}$  und  $\mathbf{R}$  miteinander stehen.

**MCAR**

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow P(\mathbf{R} = 0 | \psi)$$

Werte in  $\mathbf{Y}$  fehlen “zufällig” (unabhängig von Werten von  $\mathbf{Y}$ ). Das Fehlen von Werten ist nur bestimmt durch die allgemeine Wahrscheinlichkeit, dass Werte fehlen (im Datensatz gab es eher viele oder wenige Missings).

**Beispiel**

Das Fehlen der Variable “Alter” ist weder von der Variable “Neurotizismus”, noch vom Alter der Person selbst abhängig.

**MAR**

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \psi)$$

Werte in  $\mathbf{Y}$  fehlen abhängig von  $\psi$  **und** **observierter** Information  $\mathbf{Y}_{\text{obs}}$ .

**Beispiel**

Personen mit höheren Neurotizismus-Werten geben ihr Alter seltener an als Personen mit niedrigem Neurotizismus, unabhängig ihres Alters.

## MNAR

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow ?$$

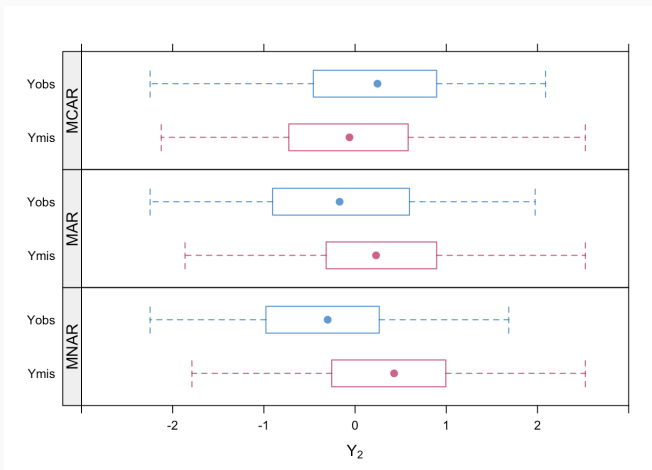
*Die Formel lässt sich nicht “vereinfachen”!*

Werte in  $\mathbf{Y}$  fehlen abhängig von  $\psi$ , observierter Information  $\mathbf{Y}_{\text{obs}}$ , **und** der **unobservierten Information**  $\mathbf{Y}_{\text{mis}}$  selbst.

### Beispiel

Das Fehlen der Alters-Variable hängt (u.A.) vom Alter der Person selbst ab. Eventuell geben ältere Personen mit hohem Neurotizismus ihr Alter seltener an. **Aber das wissen wir nicht**, da die Daten fehlen!

## Observed and unobserved values under three Missing Data-Annahmen



aus Van Buuren (2018), Kapitel 2.2.4



## Implikationen

- **MCAR:** Da die Werte zufällig fehlen, ist beim Einsatz passender statistischer Verfahren keine Verzerrung der Ergebnisse zu erwarten. Es besteht kein systematischer **Bias**; nur ein Verlust der statistischen Power durch den Datenverlust.
- **MAR:** Bei der Schätzung von Parametern ergeben sich evtl. Verzerrungen (Bias), wenn die abhängigen Variablen nicht im Modell berücksichtigt werden. Werden die abhängigen Variablen adäquat mit einbezogen, können Verzerrungen vermieden werden.

- **MNAR:** Bei der Schätzung von Parametern ergeben sich Verzerrungen. Da das Fehlen von Werten von den fehlenden Werten selbst abhängig ist, können nur “best guesses” hinsichtlich des **zugrundeliegenden Dropout-Mechanismus** getroffen werden. Ob diese korrekt sind, kann nicht empirisch überprüft werden.

Beim Vorliegen von MNAR spricht man von **"nonignorable missing data"** (Little & Rubin, 2019). Dies bedeutet, dass nicht von der gemeinsamen (Posterior-) **Verteilung der observierten Daten auf die der fehlenden Daten geschlossen werden kann**:

$$P(\mathbf{Y}|\mathbf{Y}_{\text{obs}}, \mathbf{R} = 1) \neq P(\mathbf{Y}|\mathbf{Y}_{\text{obs}}, \mathbf{R} = 0)$$

Das impliziert, dass eine Schätzung (Imputation) auf Basis der vorliegenden Werte nicht ohne weiteres möglich ist. Es müssen Annahmen getroffen werden, die **"über die Daten hinaus gehen"**.

## Was bedeutet das für die Analyse von RCT-Daten?

### MCAR

- Complete Case-Analysen und andere Ad Hoc-Verfahren führen nicht zu einer systematischen Verzerrung der Ergebnisse (aber durchaus zu einem Verlust statistischer Power/“Effizienz”).
- Die MCAR-Annahme ist typischerweise für RCT-Daten **nicht sehr plausibel**. (Bell et al., 2014; Mallinckrodt et al., 2004)
- Selbst wenn die MCAR-Annahme zutrifft, können Verfahren wie Multiple Imputation genutzt werden, z.B. um Konfidenzintervalle korrekter zu schätzen. (vgl. Pedersen et al., 2017)

## Was bedeutet das für die Analyse von RCT-Daten?

### MAR

- Complete Case-Analysen führen zu einer Verzerrung der Ergebnisse.
- Verfahren wie Multiple Imputation, Full Information Maximum Likelihood (FIML) oder Mixed-Effect Models (eingeschränkt) können aber **genutzt werden, um die MAR-Annahme abzubilden**.
- Werden diese Modelle korrekt angewandt, vermeidet dies gebiaste Ergebnisse und führt zu einem korrekten Miteinbezug der Unsicherheit durch fehlende Werte (→ passende Konfidenzintervalle).

## Was bedeutet das für die Analyse von RCT-Daten?

### MNAR

- Auch Verfahren wie Multiple Imputation (denen nur die observierten Daten zugrunde liegen) können zu **Verzerrungen in den Ergebnissen führen**.
- Methoden wie **Pattern-Mixture/Selektionmodelle** oder **referenzbasierte Imputation** können genutzt werden, um die Ergebnisse unter Annahme bestimmter Dropout-Mechanismen zu analysieren. (Little & Rubin, 2019, Kapitel 15; Carpenter et al., 2013; Heckman, 1976)
- Diese Annahmen sind jedoch nicht direkt empirisch nachweisbar.

*“MNAR models are [...] typically **highly dependent on untestable and often implicit assumptions** regarding the distribution of the unobserved measurements given the observed measurements.”*

—Molenberghs et al. (2004), S. 447

### Was soll ich für meinen Trial annehmen?

*“[W]e recommend that in trials [...], all data should be used in an analysis that makes a plausible assumption about missing data. Usually this will be a MAR assumption.”*

—Bell et al. (2014)

*“The assumption of ignorability is often sensible in practice, and generally provides a natural starting point.”*

—Van Buuren (2018), Kapitel 2.2.6



- Dropout & Fehlende Daten sind bei RCTs **kaum zu vermeiden**.
- Das Fehlen von Daten kann man sich als **“Produkt” eines wahrscheinlichkeitsbasierten Prozesses** vorstellen, der von einem (realen oder angenommenen) **Missing Data-Mechanismus** gesteuert wird.
- Derartige Mechanismen lassen sich in drei “Archetypen” zusammenfassen: **MCAR, MAR** und **MNAR**.
- Wird der zugrundeliegende Missing Data-Mechanismus nicht berücksichtigt, kann dies bei der Analyse zu **Verzerrungen** und **falschen Schlußfolgerungen** führen.



- Ob Daten MAR oder MNAR sind, kann **nicht anhand der Daten selbst bestimmt werden**; beide Annahmen sind immer nur mehr oder weniger plausibel.
- Bei RCTs ist die **Annahme von MAR häufig ein guter Startpunkt**; die Ergebnisse im Fall von nonignorable missing data (MNAR) können dann z.B. durch Sensitivitätsanalysen geprüft werden (Bell et al., 2014).
- Insbesondere **multiple Imputationsverfahren** sind ein gutes Mittel, Verzerrungen der Ergebnisse durch Dropout unter Annahme von MAR vorzubeugen.



## Kurze Anmerkung zu Little's MCAR Test

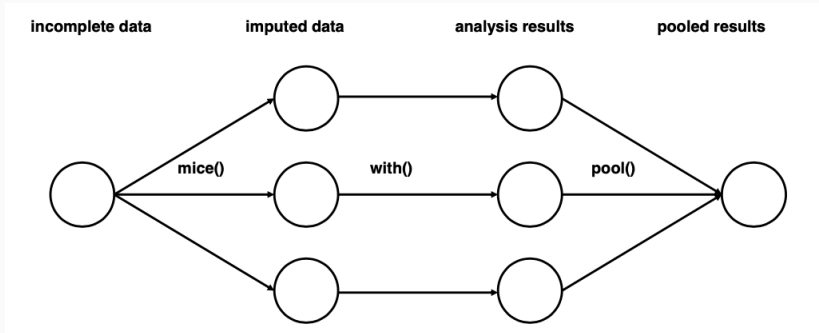
- Der  $\chi^2$  MCAR Test von Little findet immer noch häufig bei der Analyse von RCT Verwendung.
- Die Nullhypothese ist hierbei, dass Missings in den Daten zufällig auftreten ( $\rightarrow$  MCAR).
- Der **praktische Nutzen dieses Tests ist häufig begrenzt**: die Fähigkeit des Tests, die Nullhypothese (MCAR) zu verwerfen, hängt von der Größe des Datensets, und damit der statistischen Power ab. Bei großen Datensets können schon kleine Abweichungen zu einem Wert  $p < 0.05$  führen.
- Ein signifikanter MCAR-Test sagt nichts darüber, ob die Daten MAR oder MNAR sind!
- **Tipp**: stattdessen MAR als plausiblen "first start" annehmen, MNAR-Auswirkungen ggf. durch Sensitivitätsanalysen explorieren.

siehe Little (1988).

## Multiple Imputation

---

- Multiple Imputation (**MI**, Rubin, 1987) ist eine der flexibelsten und **gängigsten Methoden** zum Umgang mit fehlenden Werten.
- **Ziel von MI:** auf Basis der Verteilung der beobachteten Daten plausible Werte für die fehlenden Werte schätzen (“imputieren”).
- Um die Unsicherheit bei der Schätzung fehlender Werte abzubilden, werden **mehrere (“multiple”) Imputationen** für jeden fehlenden Datenpunkt erzeugt.
- Die so generierten vollständigen Datensätze werden dann **simultan analysiert** (z.B. Berechnung des Stichprobenmittelwerts) und Ergebnisse abschließend **gepoolt**.



aus Van Buuren & Groothuis-Oudshoorn (2011).

- MI kann unter **Annahme von MCAR und MAR** (sowie unter bestimmten Voraussetzungen auch MNAR) verwendet werden.
- Wenn korrekt angewandt führt MI zur **verzerrungsfreien Schätzung von Populationsparametern** (Regressionsgewichte, Populationsmittelwerte, Korrelationen, etc.) sowie deren Varianz (“asymptotically unbiased”) — trotz des Vorliegens fehlender Werte.

The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing data. Royston

-> Unbiased + Effizient

MICE: FCS oder Sequential Regression Multivariate Imputation

Prädiktoren der unvollständigen Variablen einbezogen

(Auxiliarvariablen): - Reduce bias, make MAR more plausible - Improve imputations, reduce variance

the imputation model should include every variable that both predicts the incomplete variable and predicts whether the incomplete variable is missing.

Thus, justification of the MICE procedure has rested on empirical studies [10, 49] rather than theoretical arguments.

MI, like all missing data procedures that are based around an MAR





$$\begin{aligned} T &= \overbrace{\left( \frac{1}{m} \sum_{\ell=1}^m \bar{U}_{\ell} \right)}^{\bar{U}} + \left( 1 + \frac{1}{m} \right) \overbrace{\left( \frac{1}{m-1} \sum_{\ell=1}^m (\hat{Q}_{\ell} - \bar{Q})(\hat{Q}_{\ell} - \bar{Q})' \right)}^B \\ &= \bar{U} + \left( 1 + \frac{1}{m} \right) B \\ &\Rightarrow \bar{U} + B \quad \text{as } m \rightarrow \infty \end{aligned}$$

### Bulleted Lists

- Element A
- Element B
  - B.1
  - B.2
- Element C

## Elements

---

The theme provides sensible defaults to  
`\emph{emphasize}` text, `\alert{accent}` parts  
 or show `\textbf{bold}` results.

In Markdown, you can also use `_emphasize_` and `**bold**`.

becomes

The theme provides sensible defaults to *emphasize* text, **accent** parts or  
 show **bold** results (Arendt, 1989; Bandersson & Cuijpers, 2009; Smit et al., 2006)

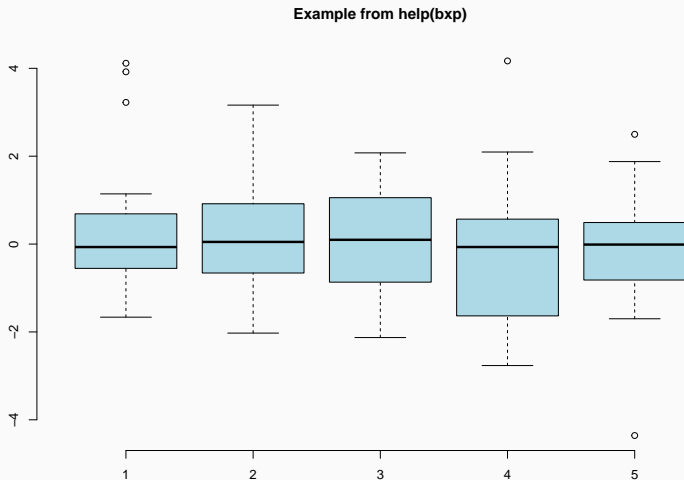
.

In Markdown, you can also use *emphasize* and **bold**.

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

The following code generates the plot on the next slide (taken from `help(bxp)` and modified slightly):

```
library(stats)
set.seed(753)
bx.p <- boxplot(split(rt(100, 4),
                      gl(5, 20)), plot=FALSE)
bxp(bx.p, notch = FALSE, boxfill = "lightblue",
     frame = FALSE, outl = TRUE,
     main = "Example from help(bxp)")
```





A simple `knitr::kable` example:

```
knitr::kable(mtcars[1:5, 1:8],  
             caption="(Parts of) the mtcars dataset")
```

**Table 1:** (Parts of) the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0

As one example of falling back into  $\LaTeX$ , consider the example of three different block environments are pre-defined and may be styled with an optional background color.

**Default**

Block content.

**Alert**

Block content.

**Example**

Block content.

## For more information:

- See the [Metropolis repository](#) for more on Metropolis
- See the [RMarkdown repository](#) for more on RMarkdown
- See the [binb repository](#) for more on binb
- See the [binb vignettes](#) for more examples.

- Akl, E. A., Shawwa, K., Kahale, L. A., Agoritsas, T., Brignardello-Petersen, R., Busse, J. W., Carrasco-Labra, A., Ebrahim, S., Johnston, B. C., Neumann, I., & others. (2015). Reporting missing participant data in randomised trials: Systematic survey of the methodological literature and a proposed guide. *BMJ Open*, 5(12), e008431.
- Arendt, H. (1989). *Vom leben des geistes, band 2, das wollen. Aus dem amerikanischen von hermann vetter*. Piper.
- Bandersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: A meta-analysis. *Cognitive Behaviour Therapy*, 38(4), 196–205.
- Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C.-H. (2014). Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, 14(1), 1–8.
- Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6), 1352–1371.
- Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st ed.). Chapman & Hall/CRC Press.

- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement*, volume 5, number 4 (pp. 475–492). NBER.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- Mallinckrodt, C. H., Watkin, J. G., Molenberghs, G., & Carroll, R. J. (2004). Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 3(3), 161–169.
- Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3), 445–464.
- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Theory of statistics* (pp. 697–716). University of California Press.
- Palmer, M. J., Mercieca-Bebber, R., King, M., Calvert, M., Richardson, H., & Brundage, M. (2018). A systematic review and development of a classification framework for factors associated with missing patient-reported outcome data. *Clinical Trials*, 15(1), 95–106.

- Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9, 157.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials*, 11(1), 1–8.
- Smit, H., Cuijpers, P., Oostenbrink, J., Batelaan, N., Graaf, R. de, & Beekman, A. (2006). Costs of nine common mental disorders: Implications for curative and preventive psychiatry. *The Journal of Mental Health Policy and Economics*, 9(4), 193–200.
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Bmj*, 338.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(1), 1–67.

- Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Bulletin of the World Health Organization*, 85, 867–872.
- Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4), 368–376.