Addressing complex missing data issues: A case study

Joint work with Donald B. Rubin, Roee Gutman

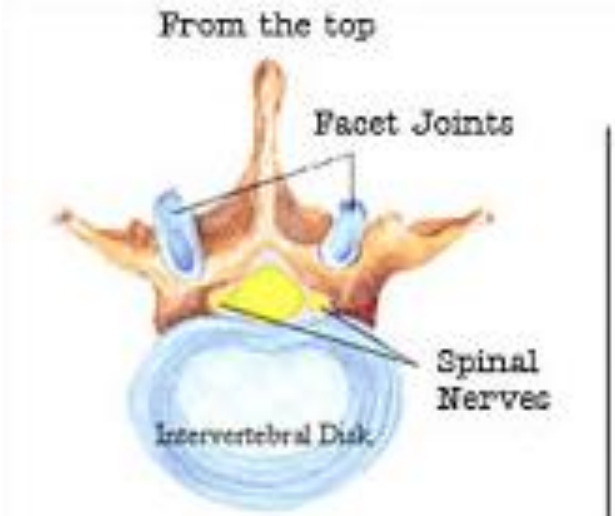
Victoria Liublinska

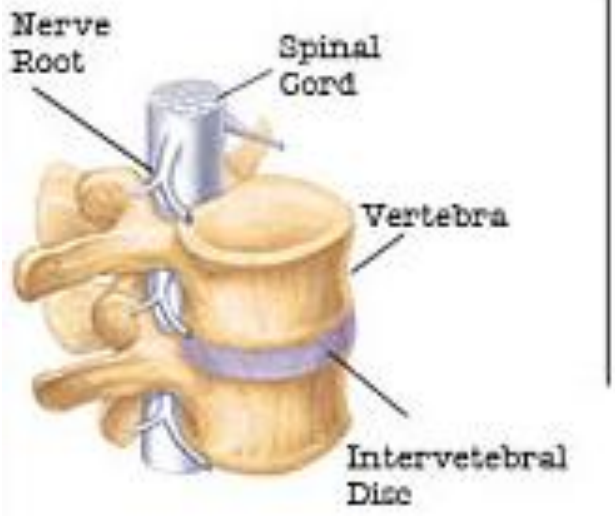
Harvard Statistics Department

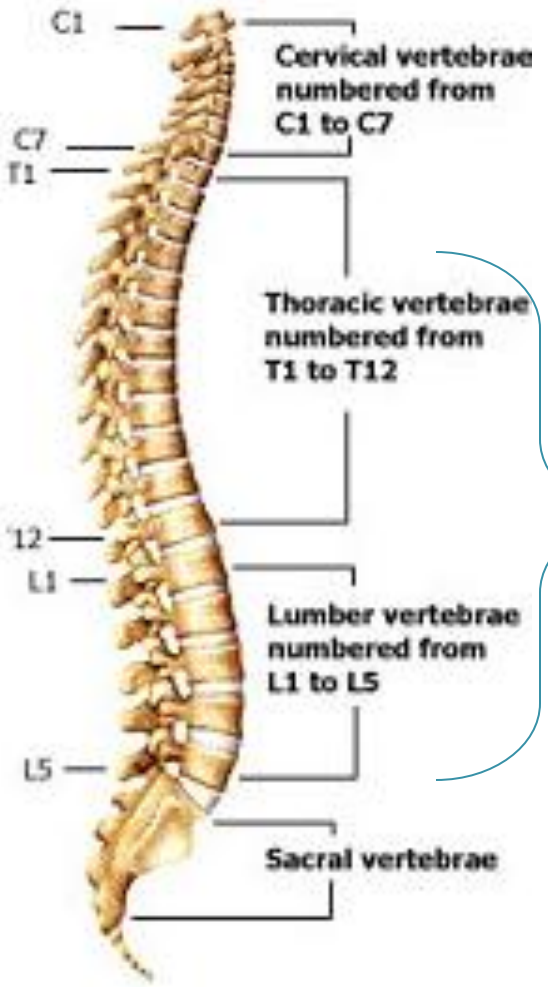
Introduction

- Background
- Objective: New Kyphoplasty System (NKS) vs. traditional vertebroplasty for the treatment of *vertebral compression fractures*.

Introduction: Vertebroplasty





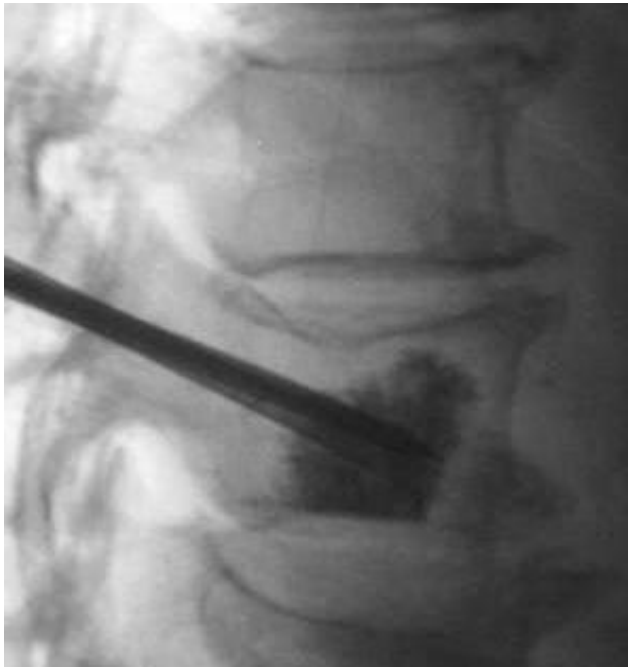


T4 - L5

(up to 3 locations)

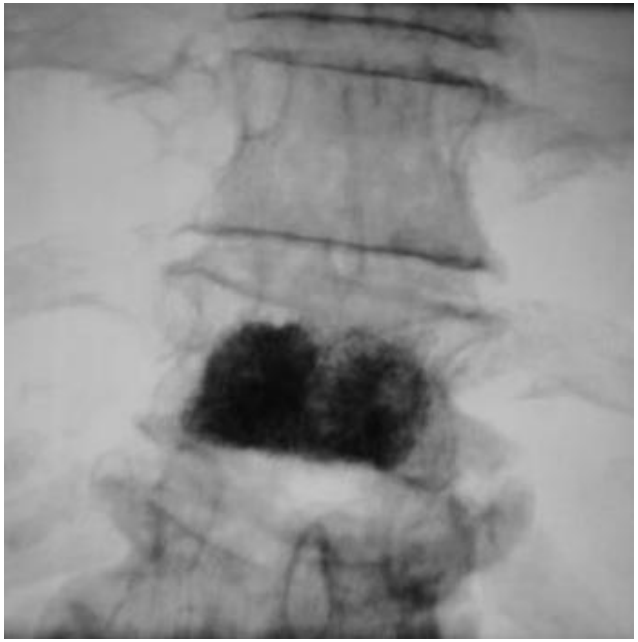
- Vertebroplasty procedure

Introduction: Vertebroplasty



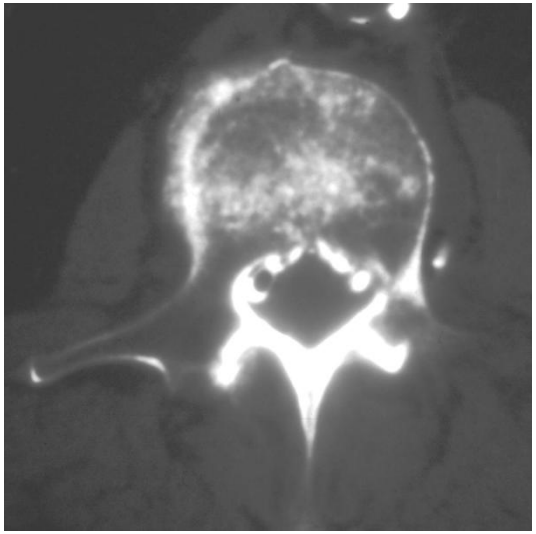
Ideal cement distribution remains in the anterior two-thirds of the vertebral body.

Final results:

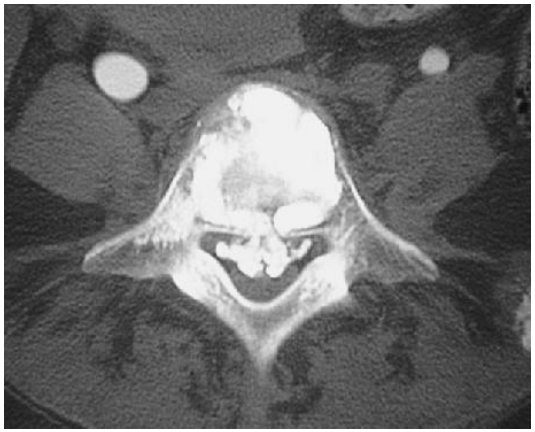


Introduction: Complications

Small Cement Leak



Moderate Cement Leaks



Large Cement Leak

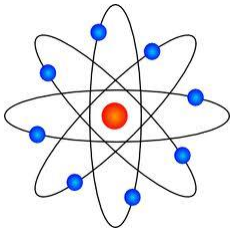
Paraplegia

Introduction: NKS device

The study tested a new device (NKS) that partially restricts the flow of bone cement and can potentially reduce leaks.

Steps to Solve a Complex Problem

1) Scientific aspect



2) Study design

3) Break down the problem into logical steps. For each step:

A. Choose a method;

B. Acknowledge/check assumptions.

4) Conclusion.

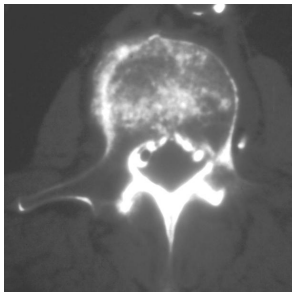
Selection of subjects and randomization

- Germany, between 3/7/08-9/17/09
- Study design:
  - Prospective
  - Open-label
  - Randomized: NKS vs. traditional procedure
- 84 subjects were evaluated, qualified, consented and randomized;
- 2:1 (NKS:control) randomization scheme;
- A total of 49 subjects were treated with NKS and 28 with control (7 were excluded from the treatment group);

Covariates

- Number of vertebral compression fractures (1,2 or 3)
- Demographics
  - age (50 or older)
  - gender
  - height, weight, BMI
  - physical activity level (minimal, light, moderate or high)
  - smoking status (never, prior, or present)
- Visual Analog Scale (VAS) pain score (6 to 10)
- Oswestry Disability Index (ODI) (0 (best) to 100 (worst))
- Duration of symptoms (in weeks)
- Center

Primary Endpoints



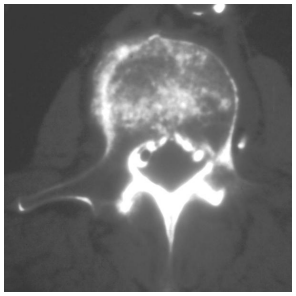
The number of cement leaks *per patient** (24h after surgery):

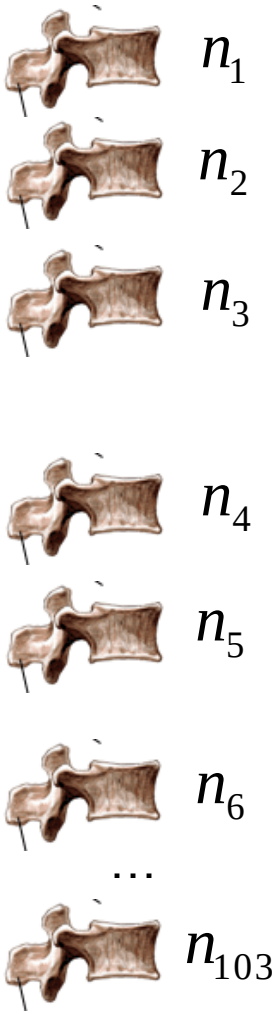
- Total number of leaks
- Number of each leak by type (B,C,S)

*Note that each patient had 1 to 3 vertebral compression fractures and each vertebrae could have multiple leaks.

There were no missing data for the primary endpoint.

Primary Endpoints





Lesson:
Randomization units = Analysis units

Collected Outcome Data

- Post-operative assessments (24h):
  - Cement Leakage
  - VAS pain score
- Assessment between discharge and three months
  - Pain (VAS) score
  - Disability (ODI) score
  - Adverse events
- Assessment between three and twelve months
  - Pain (VAS) score
  - Disability (ODI) score
  - Adverse events

Secondary Endpoints

Endpoint	Post-treatment (at 24h)	Within 3 months	Between 3 and 12 months
Average number of **Adverse Events** per patient		X	X*
Average VAS pain score	X	X	X
Average disability / quality of life (measured by ODI)		X	X

Adverse event types (6 in total):

- Adjacent Level Fracture (symptomatic / asymptomatic)
- Distant Level Fracture (symptomatic / asymptomatic)
- Re-treatment (including re-fracture)
- *Death (12-month values include deaths within 3 months)

Secondary Endpoints: Observed Rate

Endpoint	Observed rate at three months, %	Observed rate between three and twelve months, %	Rate observed in the literature, %
Re-fracture	1.3	0	2.1-2.4
Re-treatment	1.3	0	
Symptomatic Adjacent Level Fracture	2.6	1.3	8.2-15.2
Asymptomatic Adjacent Fracture	0	0	
Symptomatic Distant Level Fracture	2.6	1.3	9.8-11.6
Asymptomatic Distant Fracture	0	2.6	
Death	2.6	10.4*	11.3

*Twelve-month values include deaths within 3 months

Issues with the study

- Incorrect analysis
- Missing data in some covariates
- Missing data in secondary endpoints

Additional complications:

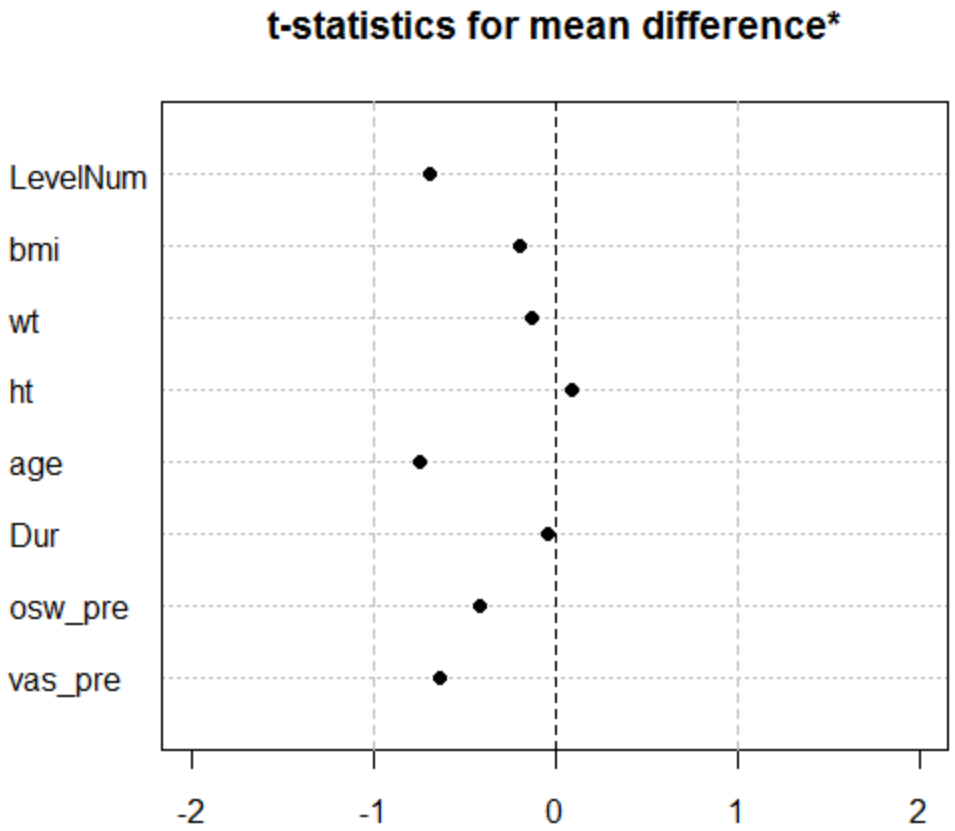
- Panel data (non-monotone missingness)
- Rare binary events
- Small N

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

— Sir Ronald Aylmer Fisher

1. ASSESSING THE BALANCE IN COVARIATES BETWEEN TREATED AND CONTROL GROUPS

Love Plot: Non-binary Covariates



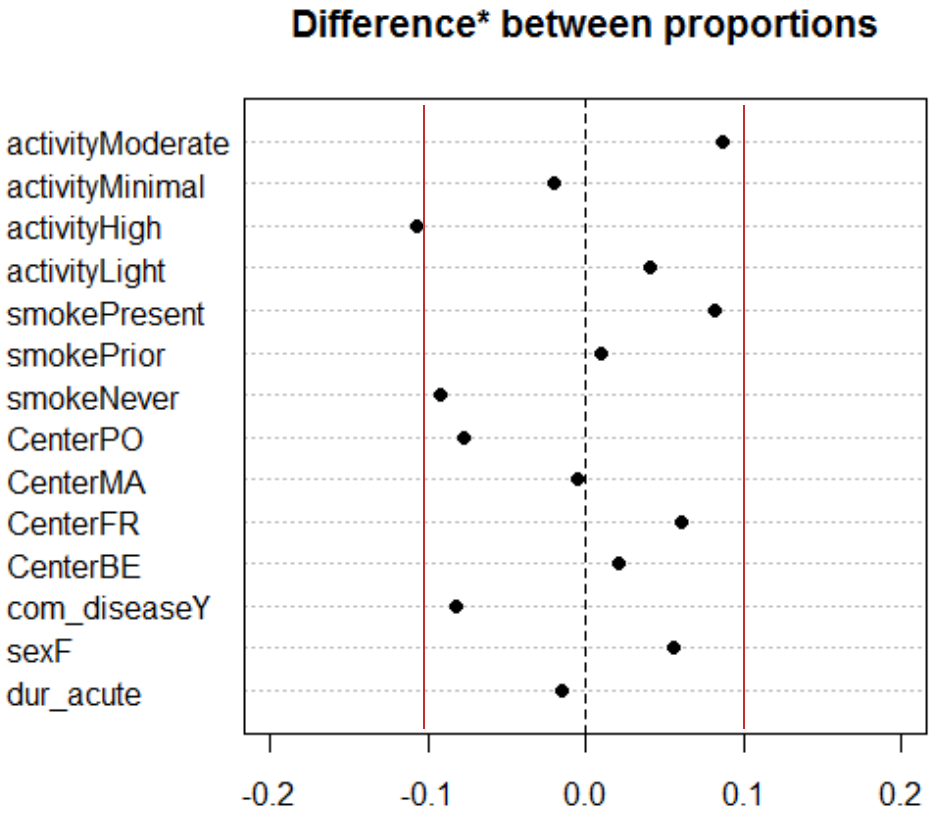
$$t = \frac{(\bar{X}_t - \bar{X}_c)}{SE}$$

$$SE = \sqrt{\frac{S_t^2}{N_t} + \frac{S_c^2}{N_c}}$$

*Treatment value minus control value

Ahmed, Husian, Love et al, (2006) Eur Heart J

Love Plot: Binary Covariates



*Treatment value minus control value

Note that initial comparison can be also made using estimated propensity scores.

2. ADDRESSING MISSING DATA ISSUES FOR COVARIATES

Missing Data in Covariates

Treated group (N = 49):

patid	tx	lev01	lev02	lev03	surgdt	vas_pre	osw_pre	psi_pre	dur_acute	dur_hospitme	dur_long	dur_time	age	sex	ht	wt	bmi	activity	race	smoke
MA-04	S	L1			4/16/2008	8	64	1	1	0			77	f	157	65	26.4	moderate	white	present
MA-05	S	L1			4/17/2008	7.5	94	1					68	f	159	64	25.3	high	white	prior
BE-10	S	L1	L2	L3	5/27/2008	8.7	70	1	0	1	12	0	73	f	165	73	26.8	light	white	

Control group (N = 28):

patid	tx	lev01	lev02	lev03	surgdt	vas_pre	osw_pre	psi_pre	dur_acute	dur_hospitme	dur_long	dur_time	age	sex	ht	wt	bmi	activity	race	smoke
BE-17	V	L1			7/23/2008	8.8	82	1	1		0		54	m				light	white	

Multiple Imputation Procedure for Covariates

- Combine treatment and control groups
  - **Remove outcome data**

- Method: Multivariate Imputation by Chained Equations (MICE) (van Buuren and Oudshoorn (2000))
  - Fully conditional speciation (FCS) (van Buuren, 2007)
  - Partially incompatible MCMC (Rubin 2003).

- 100 covariate datasets were completed be imputation.

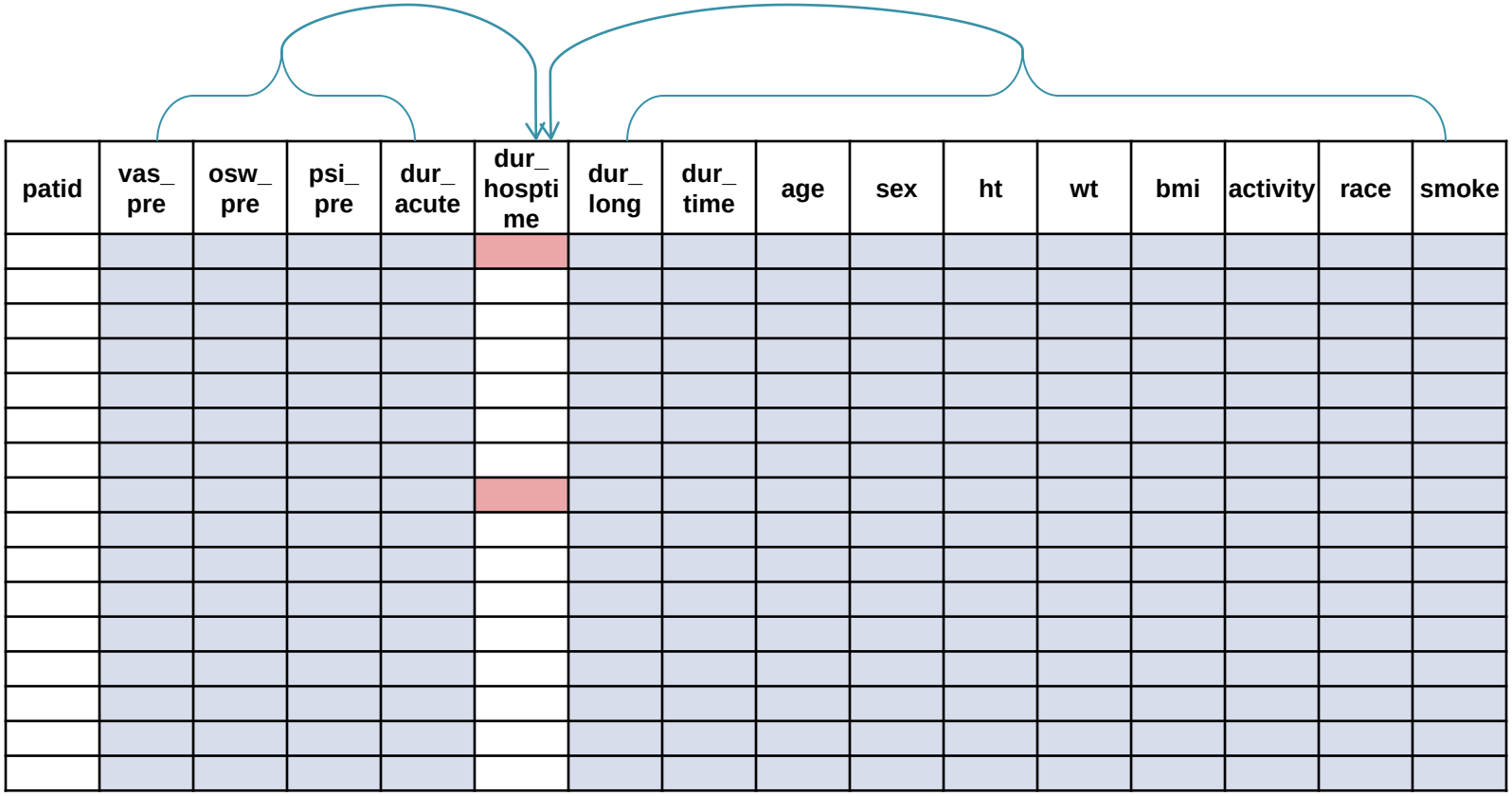
Multivariate Imputation by Chained Equations (MICE)

patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

Multivariate Imputation by Chained Equations (MICE)

patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

Multivariate Imputation by Chained Equations (MICE)

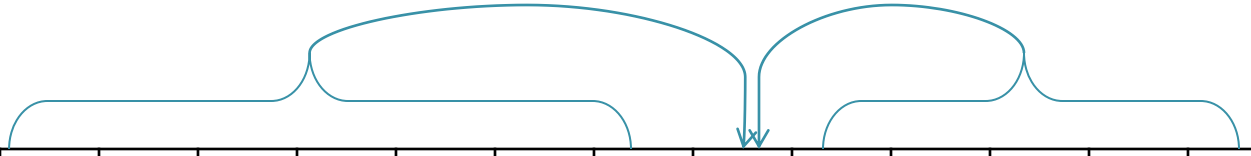


patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

Multivariate Imputation by Chained Equations (MICE)

patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

Multivariate Imputation by Chained Equations (MICE)



patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

Multivariate Imputation by Chained Equations (MICE)

patid	vas_ pre	osw_ pre	psi_ pre	dur_ acute	dur_ hospti me	dur_ long	dur_ time	age	sex	ht	wt	bmi	activity	race	smoke

- After iterating 10-15 times the procedure “converges”;
- MI’s are produced by repeating entire procedure M times.

Multivariate Imputation by Chained Equations (MICE)

- Specify conditional distributions

$$P(Y_1|Y_{-1},\theta_1)$$
$$\vdots$$
$$P(Y_p|Y_{-p},\theta_p).$$

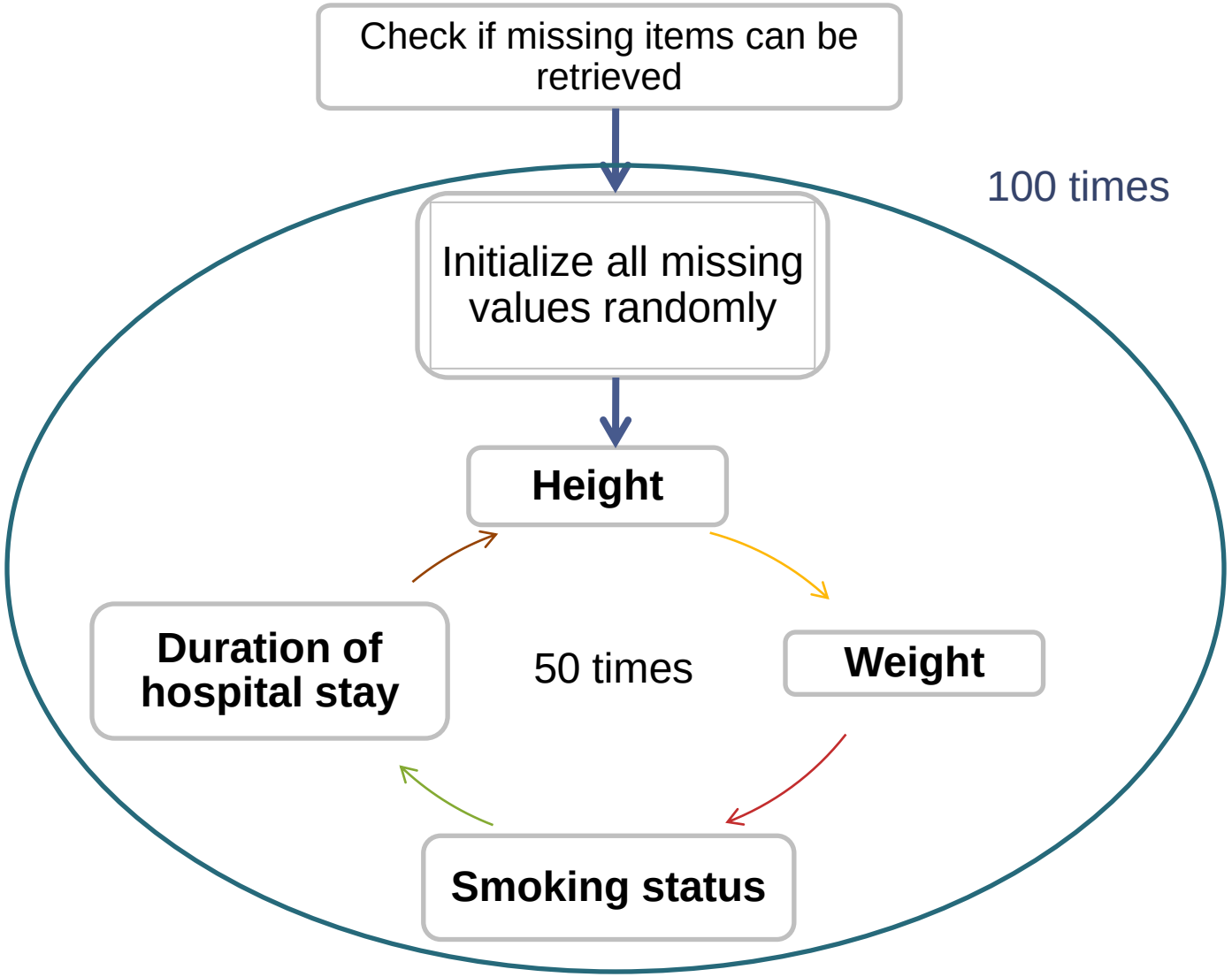
- Sequentially iterate

$$\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1|Y_1^{\text{obs}},Y_2^{(t-1)},\ldots,Y_p^{(t-1)})\\
Y_1^{*(t)} &\sim P(Y_1|Y_1^{\text{obs}},Y_2^{(t-1)},\ldots,Y_p^{(t-1)},\theta_1^{*(t)})\\
&\vdots\\
\theta_p^{*(t)} &\sim P(\theta_p|Y_p^{\text{obs}},Y_1^{(t)},\ldots,Y_{p-1}^{(t)})\\
Y_p^{*(t)} &\sim P(Y_p|Y_p^{\text{obs}},Y_1^{(t)},\ldots,Y_p^{(t)},\theta_p^{*(t)})
\end{aligned}$$

Multivariate Imputation by Chained Equations (MICE)

- Assumption: the data is missing at random (MAR)
- Caution: Conditionally specified models may be incompatible – the joint distribution may not exist.
- R-packages **mice**, **mi**

General Imputation Scheme for Covariates



3. ADDRESSING MISSING DATA ISSUES FOR SECONDARY ENDPOINTS

Secondary Endpoints

Endpoint	Post-treatment (at 24h)	Within 3 months	Between 3 and 12 months
Average number of **Adverse Events** per patient		X	X*
Average VAS pain score	X	X	X
Average disability / quality of life (measured by ODI)		X	X

Adverse event types (6 in total):

- Adjacent Level Fracture (symptomatic / asymptomatic)
- Distant Level Fracture (symptomatic / asymptomatic)
- Re-treatment (including re-fracture)
- *Death (12-month values include deaths within 3 months)

Missing data pattern for secondary endpoints: Treated group (N=49)

No	patid	tx	sex	ht	Missing data pattern		
					Post-operative	Three months	Twelve months
1	MA-05	S	f	159	Y	N	Y
2	MA-12	S	f	168	Y	N	Y
3	PO-11	S	f	170	Y	N	Y
4	BE-01	S	m	172	Y	Y	N
5	BE-22	S	f	166	Y	Y	N
6	FR-04	S	f	170	Y	Y	N
7	PO-02	S	f	154	Y	Y	N
8	BE-14	S	m	182	Y	Y	N
9	BE-03	S	f	150	Y	Y	N
10	MA-03	S	m	172	Y	Y	N
11	MA-17	S	m	179	Y	Y	N
12	PO-22	S	m	166	Y	Y	N
13	PO-16	S	f	172	Y	Y	N
14	PO-23	S	m	155	Y	Y	N
15	FR-02	S	f	157	Y	Y	N
16	BE-19	S	m	164	Y	N	N
17	FR-01	S	f	158	Y	N	N
18	FR-08	S	f	175	Y	N	N
19	FR-10	S	f	162	Y	N	N
20	MA-04	S	f	157	Y	N	N
21	MA-07	S	f	172	Y	N	N
22	MA-08	S	m	176	Y	N	N
23	MA-16	S	f	162	Y	N	N
24	MA-19	S	m	170	Y	N	N
25	BE-06	S	f	161	Y	Y	Y
26	BE-07	S	f	158	Y	Y	Y
27	BE-08	S	f	163	Y	Y	Y
28	BE-10	S	f	165	Y	Y	Y
29	BE-12	S	f	159	Y	Y	Y
30	BE-15	S	f	168	Y	Y	Y
31	BE-16	S	f	166	Y	Y	Y
32	BE-18	S	m	178	Y	Y	Y
33	BE-21	S	f	166	Y	Y	Y
34	BE-24	S	m	175	Y	Y	Y
35	FR-09	S	f	166	Y	Y	Y
36	FR-11	S	f	158	Y	Y	Y
37	PO-03	S	f	168	Y	Y	Y
38	FR-14	S	m	180	Y	Y	Y
39	MA-14	S	f	168	Y	Y	Y
40	MA-11	S	f	163	Y	Y	Y
41	MA-15	S	f	171	Y	Y	Y
42	PO-08	S	f	163	Y	Y	Y
43	FR-06	S	m	176	Y	Y	Y
44	PO-14	S	f	158	Y	Y	Y
45	PO-20	S	f	160	Y	Y	Y
46	PO-21	S	f	164	Y	Y	Y
47	PO-10	S	f	159	Y	Y	Y
48	PO-06	S	f	151	Y	Y	Y
49	FR-13	S	f	168	Y	Y	Y

Missing fraction 24% 43%

Missing data pattern for secondary endpoints: Treated group (N=49)

No	patid	tx	No. levels	age	sex	Missing data pattern		
						Post-operative	Three months	Twelve months
1	MA-05	S	1	68	f			
2	MA-12	S	1	62	f			
3	PO-11	S	2	85	f			
4	BE-01	S	1	83	m			
5	BE-03	S	1	86	f			
6	BE-14	S	1	72	m			
7	PO-02	S	1	93	f			D
8	PO-16*	S	1	70	f			
9	PO-22	S	1	82	m			D
10	PO-23	S	1	80	m			
11	MA-17	S	1	54	m			D
12	MA-03*	S	1	75	m			
13	FR-02	S	1	70	f			
14	FR-04	S	1	61	f			
15	BE-22	S	3	61	f			
16	BE-19	S	1	78	m			
17	FR-01*	S	1	85	f			
18	FR-08	S	1	56	f			
19	FR-10	S	1	77	f			
20	MA-04	S	1	77	f			
21	MA-07	S	1	84	f			D
22	MA-08	S	1	68	m			D
23	MA-16	S	1	49	f			
24	MA-19*	S	1	86	m			

* - units with small deviations from this pattern for some outcomes

D= missing due to death

Secondary endpoints "missing due to death" were treated as MAR.

Missing data pattern for secondary endpoints: Control group (N=28)

No	patid	tx	sex	ht	Missing data pattern		
					Post-operative	Three months	Twelve months
1	MA-09	V	f	154	Y	N	Y
2	MA-18	V	m	158	Y	N	Y
3	BE-02	V	f	150	Y	Y	N
4	BE-13	V	f	168	Y	Y	N
5	PO-09	V	f	162	Y	Y	N
6	PO-24	V	f	155	Y	Y	N
7	PO-26	V	f	158	Y	Y	N
8	MA-10	V	m	185	Y	Y	N
9	MA-02	V	f	163	Y	Y	N
10	MA-06	V	m	175	Y	N	N
11	BE-20	V	m	170	Y	N	N
12	PO-04	V	f	159	Y	N	N
13	BE-05	V	m	186	Y	Y	Y
14	BE-09	V	f	157	Y	Y	Y
15	BE-11	V	m	182	Y	Y	Y
16	BE-17	V	m		Y	Y	Y
17	BE-23	V	f	162	Y	Y	Y
18	FR-15	V	m	180	Y	Y	Y
19	MA-01	V	f	168	Y	Y	Y
20	MA-13	V	f	156	Y	Y	Y
21	PO-01	V	f	172	Y	Y	Y
22	PO-12	V	f	158	Y	Y	Y
23	PO-13	V	f	168	Y	Y	Y
24	PO-17	V	f	168	Y	Y	Y
25	PO-19	V	f	164	Y	Y	Y
26	FR-03	V	f	163	Y	Y	Y
27	FR-05	V	f	159	Y	Y	Y
28	FR-07	V	m	154	Y	Y	Y

Missing fraction 18% 36%

Missing data pattern for secondary endpoints: Control group (N=28)

No	patid	tx	No. levels	age	sex	Missing data pattern		
						Post-operative	Three months	Twelve months
1	MA-09	V	1	82	f			
2	MA-18	V	1	77	m			
3	PO-24	V	1	88	f			D
4	BE-02	V	2	83	f			
5	BE-13	V	1	79	f			
6	MA-02	V	1	81	f			
7	PO-09	V	1	89	f			
8	PO-26	V	1	60	f			
9	MA-06	V	1	83	m		D	D
10	PO-04	V	2	83	f		D	D
11	BE-20	V	1	76	m			

D= missing due to death

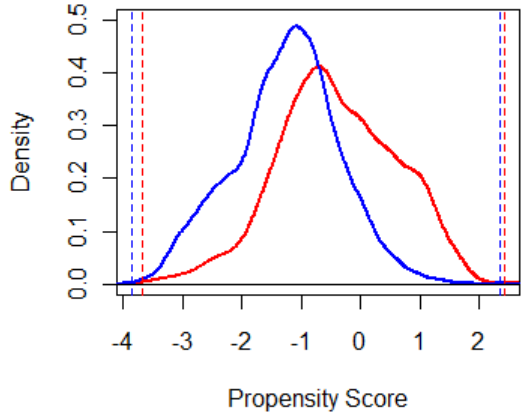
Is MAR assumption reasonable?

3.1 CHECKING OVERLAP BETWEEN RESPONDENTS AND NON-RESPONDENTS

Important: Check overlap before imputing missing data (or matching)!

- Complete overlap between the joint distributions of covariates for missing and observed units is required to avoid extrapolation.

- First step



- Ranges (including interactions), other methods will emerge from ongoing research.

Non-overlap: Respondents/Non-respondents

Observed only in control group.

At three months :

- All three male non-respondents (2 missing + 1 dead) were older than the oldest male respondent (76, 77, 83 vs. 69) ;
- Two out of three male non-respondents had lower BMI than the lowest observed in among respondents (21.5, 20 vs23.5);
- One out of two female non-respondents had “prior” smoking experience, and no female respondent was in this category;
- One male non-respondent had a duration of hospital stay longer than all male respondents;

At twelve months:

- Two female non-respondents (1 missing + 1 dead) were older than the oldest female respondent (88, 89 vs. 85);
- One male non-respondent was older than the oldest male respondent (83 vs. 77).

Note that by using responses from healthier subjects the imputation procedure produces more conservative results.

3.2 MULTIPLE IMPUTATION FOR SECONDARY ENDPOINTS

Missing-Data Imputation Procedure

- Secondary outcome data split into treated and control parts.
- Two analysts perform multiple imputation; *blinded to each other's outcome data*.
- Method for obtaining imputations: MICE (non-monotone pattern).

Missing-Data Imputation Procedure for Secondary Endpoints

Secondary endpoints that had to be imputed:

	3 months	12 months
• Pain (VAS) score	x	x
• Disability score(ODI)	x	x
Adverse Events:		
• Symptomatic Adjacent Level Fracture	x	x
• Symptomatic Distant Level Fracture	x	x
• Re-treatment	x	x
• Asymptomatic Adjacent Fracture	x	x
• Asymptomatic Distant Fracture	x	x
• Death	x	x

Missing-Data Imputation Steps

No	Covariates	ODI and VAS		Adverse Events	
		Three months	Twelve months	Three months	Twelve months
1					
2					
…	…	…	…	…	…
N					

No	Covariates	ODI and VAS		Adverse Events	
		Three months	Twelve months	Three months	Twelve months
1					
2					
…	…	…	…	…	…
N					

Missing-Data Imputation Steps

No	Covariates	ODI and VAS		Adverse Events	
		Three months	Twelve months	Three months	Twelve months
1					
2					
…	…	…	…	…	…
N					

VAS sample values: 0, 0, 0, 0.3, 0.3, 0.4, 0, 0, 3.5, 4, 5, 2, 0.8 …

- Semi-continuous distribution
  - Logistic regression to impute zero-indicator;
  - PMM to impute non-zero part.
- Same for VAS and ODI at three and twelve months.

Missing-Data Imputation Steps

No	Covariates	ODI and VAS		Adverse Events	
		Three months	Twelve months	Three months	Twelve months
1					
2					
…	…	…	…	…	…
N					

Adverse events are rare and **don't have good predictors**

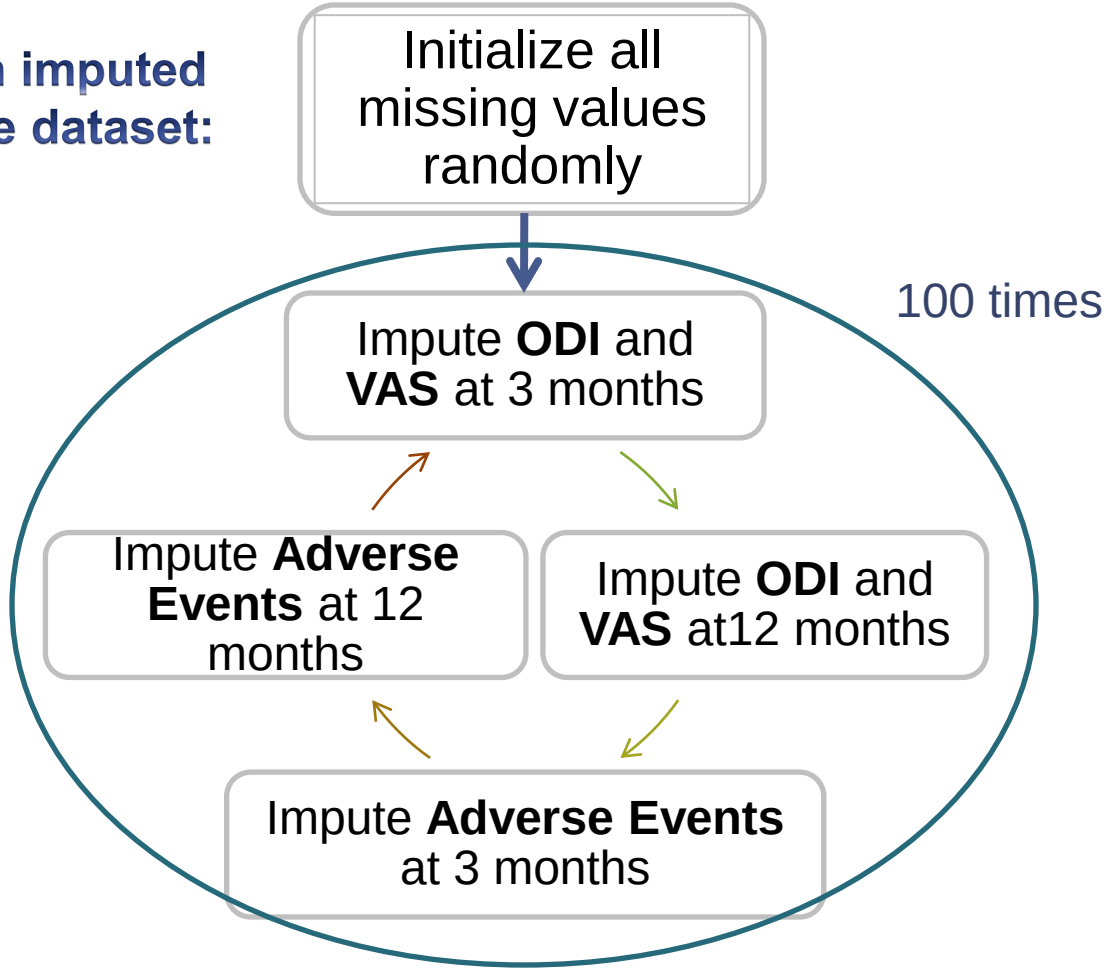
- *Adjustment Cell Method* with a "file concatenation"

		VAS		
		Zero	Below median	Above median
ODI	Zero			
	Below median			
	Above median			

- Same for Adverse Events at three and twelve months.

General Imputation Scheme: Endpoints at 3 and 12 Months

For each imputed covariate dataset:



100 datasets completed by imputation were generated.

3.3 ANALYSIS AND CONCLUSION

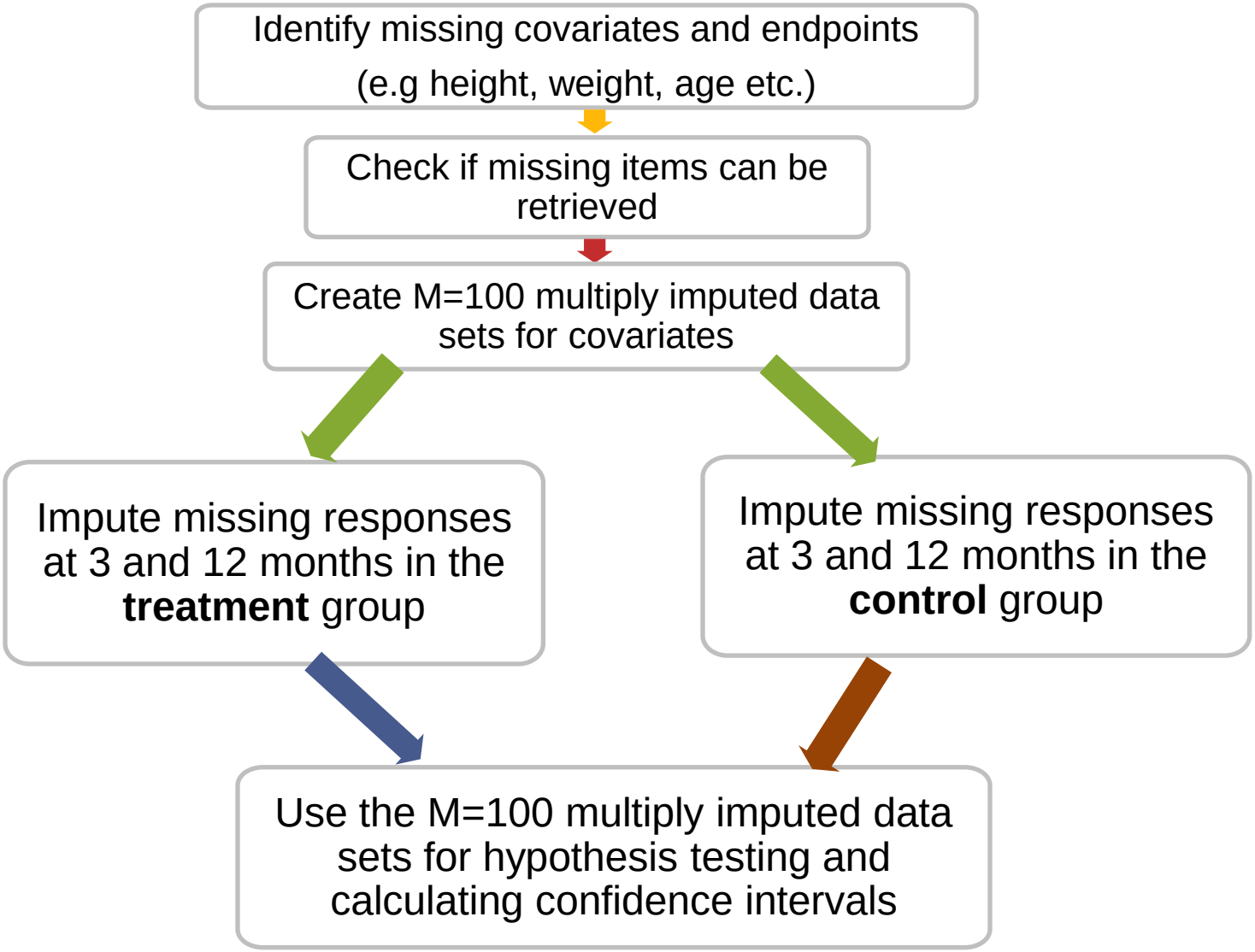
Analysis of MIed datasets

- **Fisher Randomization Test** (applied to each dataset)
- 100 *p*-values were combined using a procedure *analogous* to Rubin's Combining Rules (Rubin, 1987);
  - In 2009 Rubin proposed a simple work-around (described in C. Licht's thesis):

$$z_l = \Phi^{-1}(1 - p_l)$$

  - Combine obtained *z*-scores using the usual combining rules.
  - Note that this method can only be applied to a one sided *p*-value.

General Imputation Scheme: Summary



Results: Primary Endpoints

Endpoints	One-sided p-values via Fisher Exact for null of no difference		Asymptotic 95% Confidence intervals		
	Alternative Hypothesis: Treatment better than Control	Alternative Hypothesis: Control better than Treatment	Treatment	Control	Difference
Total **Number of Leaks** per Person	**0.0016**	0.999	(0.52, 1.20)	(1.27,2.51)	(-1.73, -0.34)
… Type B	**0.018**	0.995	(0.04, 0.29)	(0.23, 0.62)	(-0.49, -0.04)
… Type C	**0.003**	0.999	(0.16, 0.45)	(0.46, 1.11)	(-0.83, -0.13)
… Type S	0.13	0.908	(0.13, 0.64)	(0.24, 1.11)	(-0.79, 0.21)

- "**Blue**": reject, at 0.05 level, null hypothesis of no difference in favor of alternative hypothesis that treatment is better than control;
- "Black": do not reject null hypothesis of no difference.

Results: Secondary Endpoints (Events per person)

Adverse events		One-sided p-values via Fisher Exact for null of no difference	
		Alternative Hypothesis: Treatment better than Control	Alternative Hypothesis: Control better than Treatment
Results at three months	Retreatment	0.999	0.388
	Symptomatic Adjacent Fracture	0.304	0.889
	Symptomatic Distant Fracture	0.999	0.377
	Asymptomatic Adjacent Fracture	0.999	0.996
	Asymptomatic Distant Fracture	0.999	0.996
	Death	0.126	0.998
	Any event before 3 months	0.29	0.833
Results between three and twelve months	Retreatment	0.995	0.993
	Symptomatic Adjacent Fracture	0.995	0.461
	Symptomatic Distant Fracture	0.271	0.996
	Asymptomatic Adjacent Fracture	0.997	0.989
	Asymptomatic Distant Fracture	0.477	0.896
	Death**	0.593	0.679
	Any event during 3 to 12 months	0.319	0.802

** Includes deaths within three months

Conclusions

- For primary endpoints (cement leaks) NKS is superior.
  - Using randomization based analysis.
- For secondary endpoints, NKS showed equivalent results to control procedure.
  - Traditional analysis modified to deal with missing data.

Steps for careful data analysis

- Understand scientific aspect
- Learn about study design
- Check covariate balance
  - Address missing data issue in covariates
- Analyze outcome data
  - Check overlap between respondents and non-respondents
  - Address missing data issue in outcome data
  - Acknowledge and check assumptions
- Form conclusions
- Perform sensitivity analysis

Thank you