14. Rehabilitationswissenschaftliches Kolloquium
Rehabilitationsforschung in Deutschland: Stand und Perspektiven
Hannover, 28.2.-2.3.2005

Multiple Imputation in der Rehabilitation

Ersetzung von fehlenden Werten mit Multipler Imputation an einem empirischen Datensatz

Dipl.-Psych. Wilmar Igl

Universität Würzburg
Arbeitsbereich Reha-Wissenschaften

Einleitung

- Fehlende Werte als allgegenwärtiges Problem der Forschung, auch in den Rehabilitationswissenschaften

- Resultierende Probleme durch fehlende Werte:
  - Verzerrung (*bias*) der Ergebnisse
  - Verringerung der Teststärke von statistischen Verfahren

- Maximum-Likelihood-Ansatz (ML, FIML, EM) und Multiple Imputation als *state-of-the-art*-Methoden zur Behandlung von fehlenden Werten (vgl. Schafer & Graham, 2002)

I. Theoretischer Hintergrund

Missing Data Prozesse

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Non-Missing At Random (NMAR) / Informative Drop-Out (vgl. Little & Rubin, 2002)

→ Abhängig von zu Grunde liegendem Missing Data Prozess können Verzerrungen der Daten/Ergebnisse auftreten!

→ Vorbedingung für die Auswahl von Methoden zur Behandlung von Fehlwerten!

MCAR - Diagramm

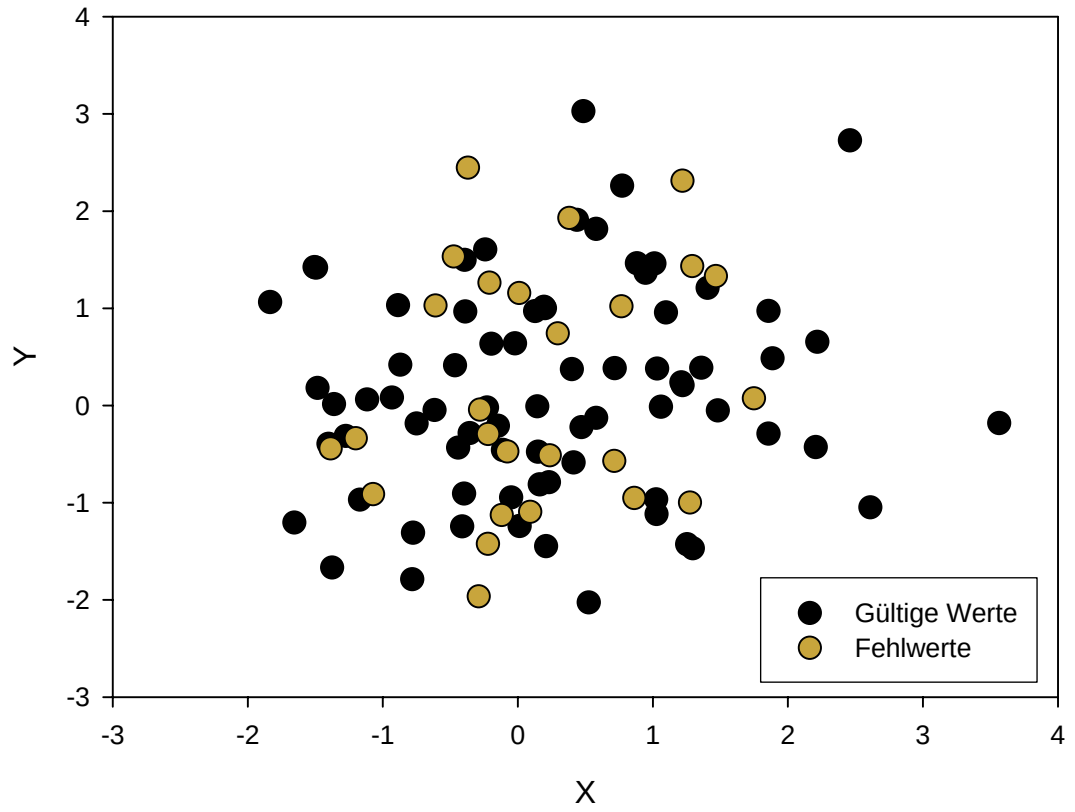


Abb. 1: Streudiagramm mit Fehlwerten (braun) unter der Bedingung MCAR

MAR - Diagramm

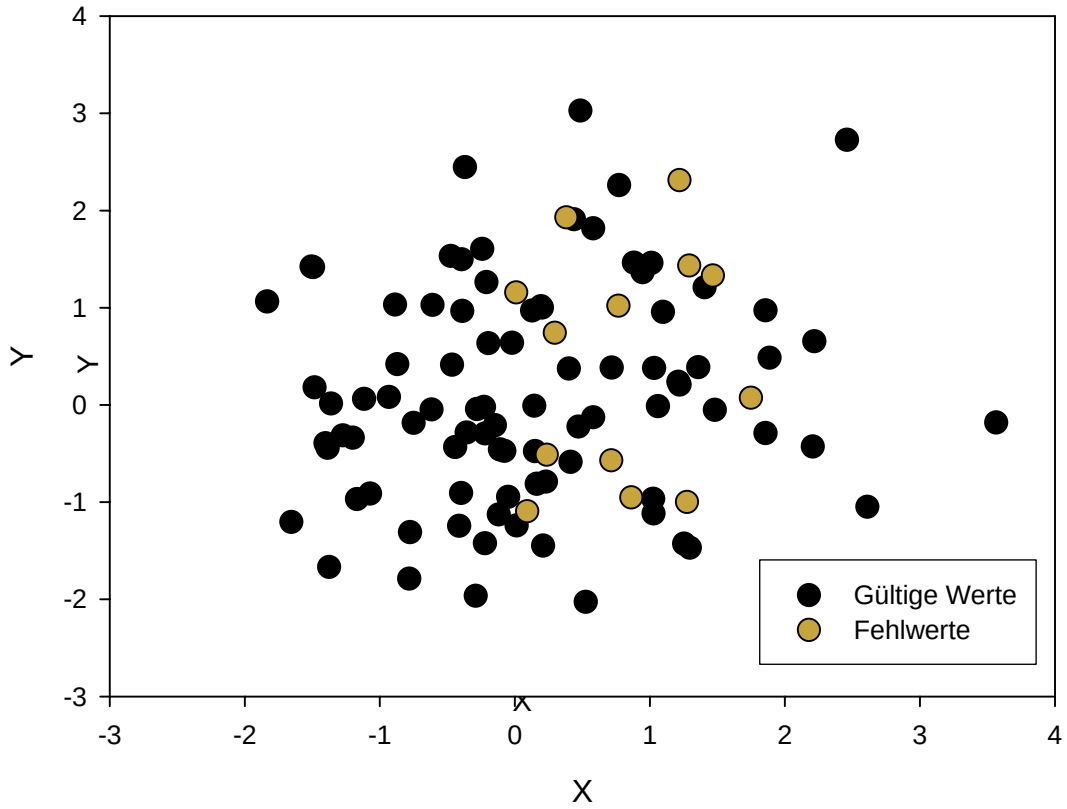


Abb. 2: Streudiagramm mit Fehlwerten (braun) unter der Bedingung MAR

NMAR - Diagramme

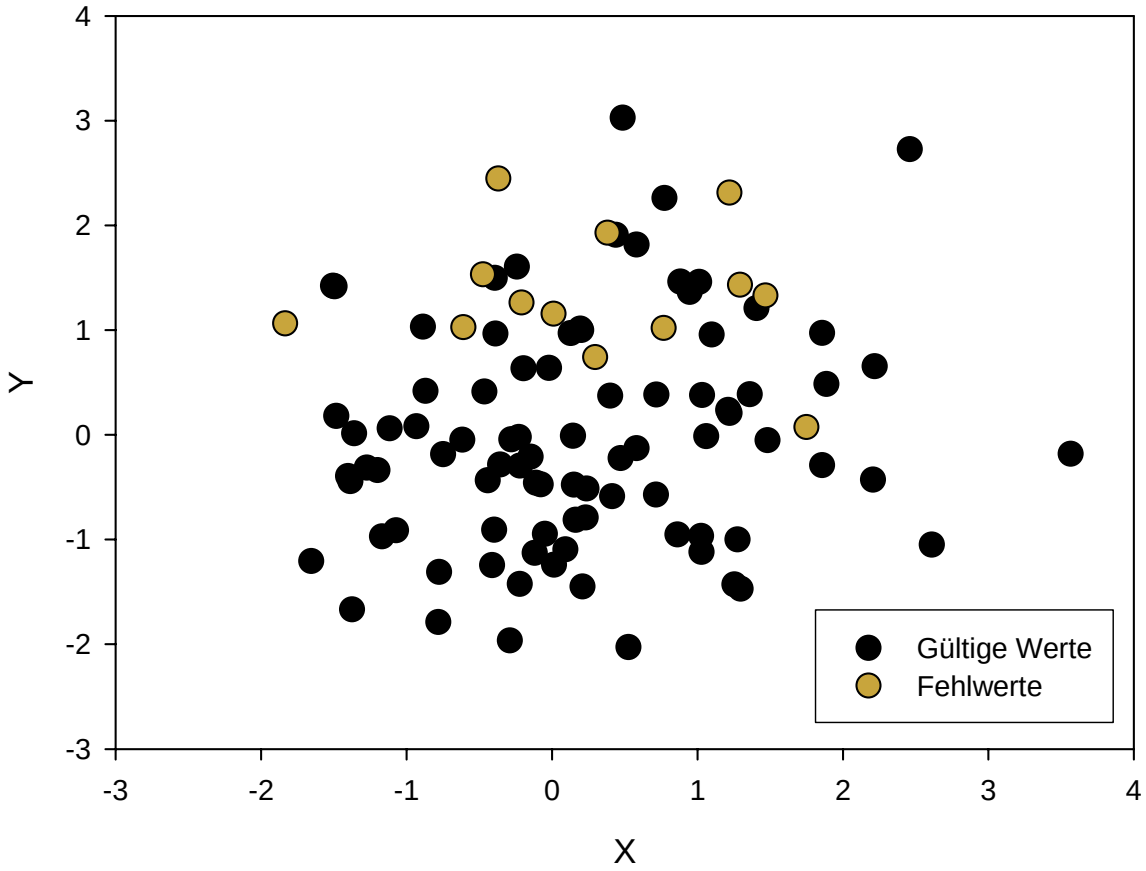


Abb. 3: Streudiagramm mit Fehlwerten (braun) unter der Bedingung NMAR

Listenweiser Fallausschluss

- **Vorgehen**: Ausschluss aller unvollständiger Fälle/ Variablen („*complete case approach*“)

- **Anwendung bei:**
  - MCAR
  - große Stichprobe
  - starke Effekte

- **Nachteile**:
  - Reduktion der Stichprobe bis zur Unbrauchbarkeit möglich
  - MCAR ist in der Forschungspraxis selten gegeben

Paarweiser Fallausschluss

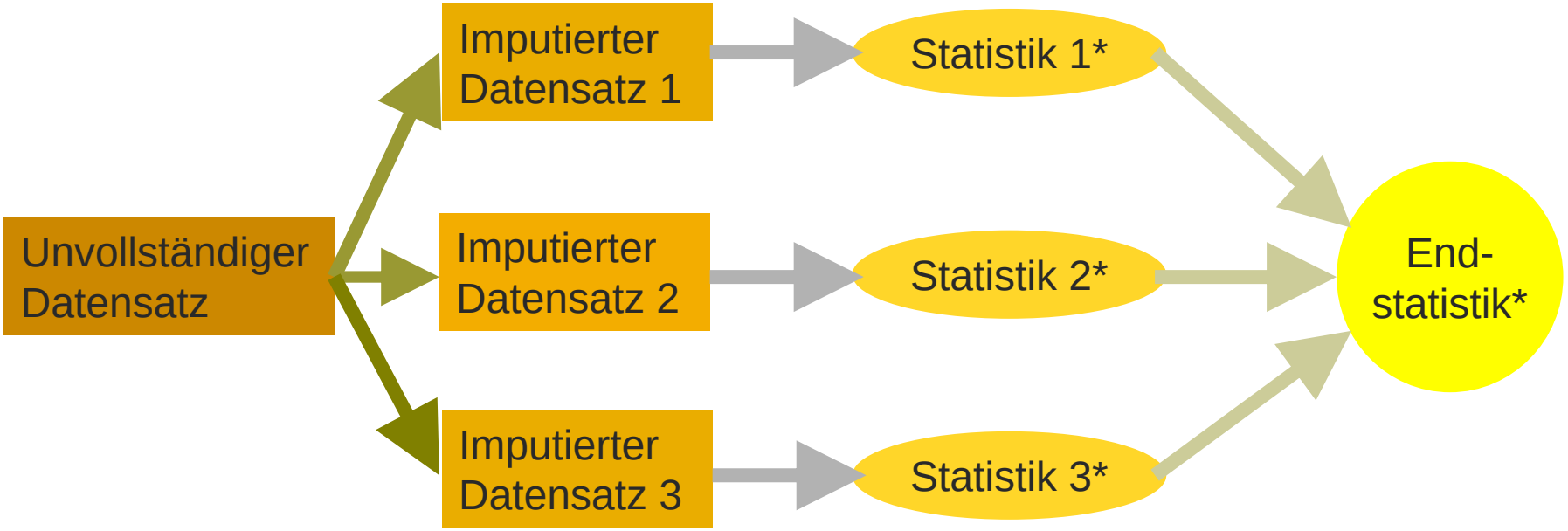
- **Vorgehen**: Alle gültigen Fälle, der in die Berechnung eingehenden Variablen, werden ausgewertet. Verteilungscharakteristika der gültigen Werte werden übernommen ("*available case approach*")
- **Anwendung bei:**
  - MCAR
  - Berechnung von Korrelationen, Mittelwerten, Streuungen
- **Nachteile:**
  - Statistiken können auf unterschiedlichen Stichproben von Beobachtungen basieren (unterschiedliches N !)
  - mathematische Inkonsistenzen möglich (z.B. zwischen Korrelationen zweier Variablen X, Y und deren Partialkorrelationen mit Z)

Multiple Imputation

- **Multiple Imputation (MI):**
  - mehrfache Ersetzung (=Imputation) von fehlenden Werten
  - Erweiterung von einfachen Imputationsmethoden (z. B. Regression,...)
- **Anwendung bei:**
  - multivariate Normalverteilung
  - MAR/ MCAR
  - große Stichprobe (>> 100)
  - akzeptabler Anteil fehlender Werte abhängig vom „Anteil fehlender Information“

Multiple Imputation: Die Grundidee

IMPUTATION ⇨ ANALYSE ⇨ INTEGRATION



*Punktschätzer und Standardfehler

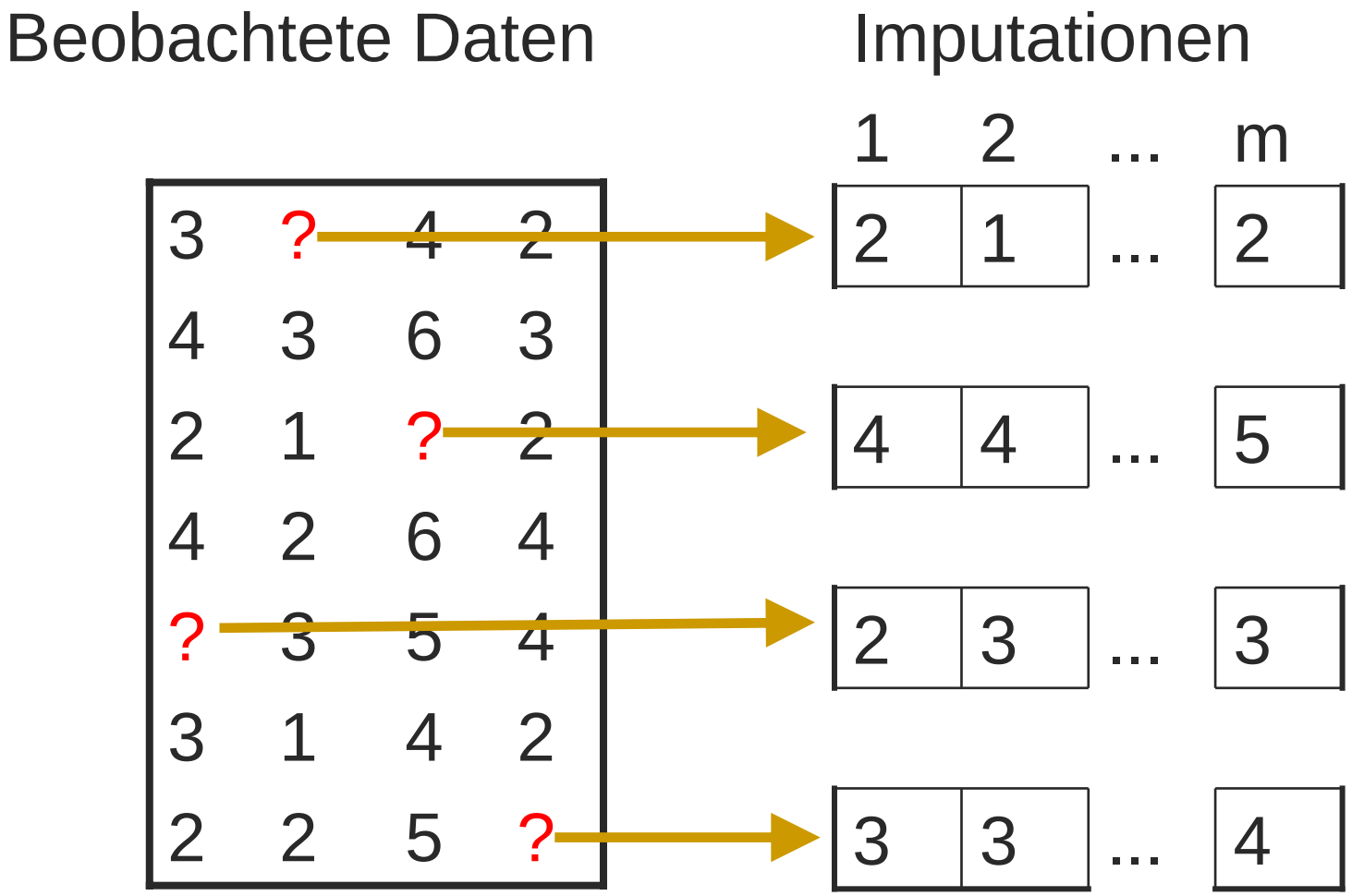
1) Imputation: DA-Algorithmus

- Iterative „Ziehung“ von Parametern (z.B. NV(μ, σ)) und Variablenwerten aus einer prädiktiven Bayes-Verteilung

$$P(Y_{mis}, \theta \mid Y_{obs})$$

- Durchführung von k Schleifen (z.B. 1000) bis der Algorithmus „konvergiert“ (bis zur Unabhängigkeit der Schätzungen)

1) Imputation: Datenbeispiel



2) Analyse

- Berechnung statistischer Parameter (Punktschätzer und ihre Standardfehler) mit Hilfe von Standard-Statistik-Software (SPSS, SAS,..)

  Beispiele: Mittelwerte, Regressionskoeffizienten, Kovarianzen und Korrelationen, ...

- Berechnung der zugehörigen Standardfehler (SE) notwendig

3) Integration (nach Rubin, 1987)

- MI Punktschätzer:
  Berechnung des arithmetischen Mittels der m Statistiken (z.B. Mittelwerte) aus m imputierten Datensätzen

- Varianz (gesamt) = Varianz (innerhalb der m Datensätze) + Varianz (zwischen den m Datensätzen)

- Berechnung von weiteren Statistiken, z. B. Freiheitsgrade, t-Werte, p-Werte, Konfidenzintervalle (95%)

Multiple Imputation

- **Vorteile:**
  - Nutzung der verfügbaren Information in beobachteten Daten
  - Steigerung der Effizienz der statistischen Auswertung aufgrund von vollständigen Datensätzen
  - Auskunft über die Unsicherheit der Ergebnisse
- **Nachteile:**
  - höherer Zeitaufwand
  - Einarbeitung notwendig

II. Methodik

Stichprobe

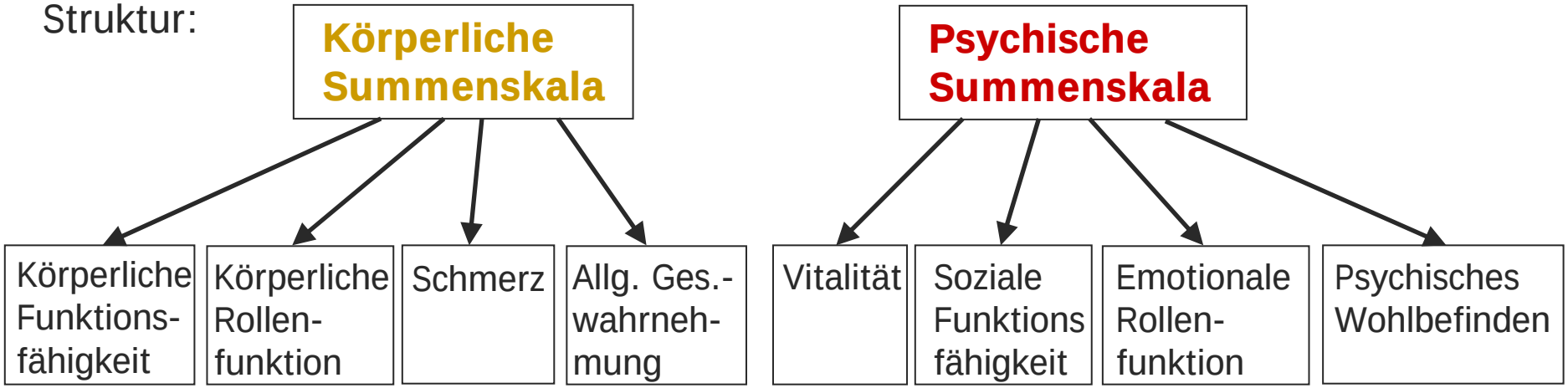
	Gesamt	Ortho/Rheuma	Kardio
N(eligible Pat.)	2012 (100%)	1164 (58%)	848 (42%)
N(teilnehm. Pat.)	**1145 (100%)**	745 (65%)	400 (35%)
Geschlecht	N(F) = 42% N(M) = 58%	N(F) = 53% N(M) = 47%	N(F) = 22% N(M) = 78%
Alter	M = 50 J. SD = 8.6 J.	M = 49 J. SD = 8.9 J.	M = 51 J. SD = 7.9 J.

Studiendesign

- Datenerhebung im Rahmen des RFB-Projekts A7 „Änderungssensitivität“ (Faller, Zwingmann & Igl)
- prospektives Ein-Gruppen-Design
- Erhebung von Daten zur gesundheitsbezogenen Lebensqualität (SF-36)
- Messzeitpunkte:

SF-36 (Bullinger & Kirchberger,1998)

Struktur:



Beispiele für Items:

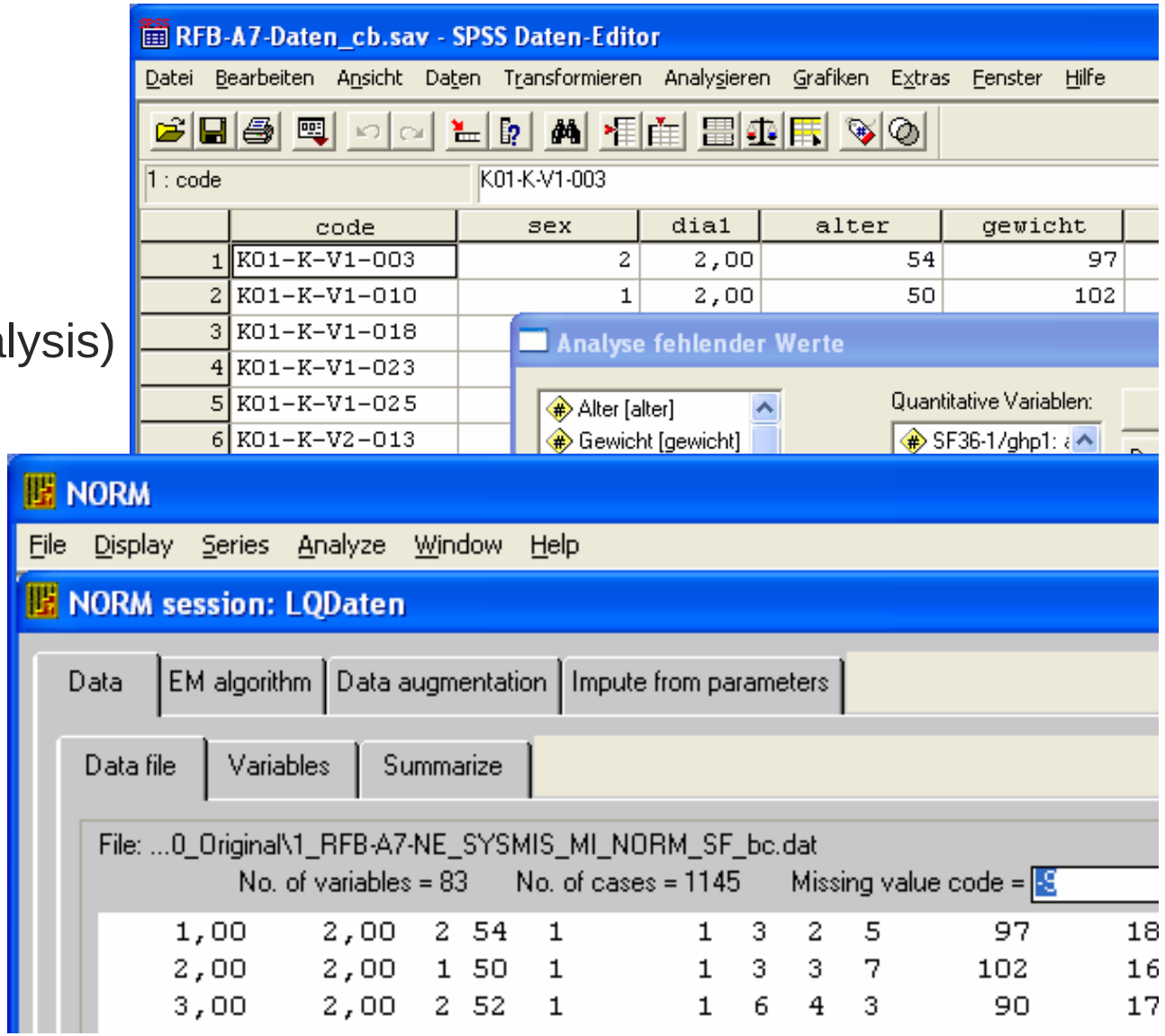
- „Treppensteigen“
- „weniger schaffen“
- „Behinderung durch Schmerz“
- „allg. Gesundheitszustand“

Beispiele für Items:

- „voller Energie sein“
- „Kontakte beeinträchtigt“
- „weniger schaffen“
- „ruhig und gelassen“

Software

- SPSS 12
- SPSS-Modul MVA
  (Missing Value Analysis)
- NORM V2.03



Missing Data Analyse

- Missing Data Analyse mit dem **SPSS-Modul MVA** und **NORM**
- **82 Variablen**
- **Anteil von Missing Data**: 20% (zwischen 16% und 37% pro Variable)

- **MCAR-Test nach Little**:
  Chi-Quadrat = 17438.3, DF = 15881, p < 0.001

⇨ Es können **Verzerrungen bei konventionellen MD-Methoden** auftreten (complete case/ available case approach)

NORM

File Display Series Analyze Window Help

File: summary.out

NUMBER OF OBSERVATIONS = 1145
NUMBER OF VARIABLES = 82

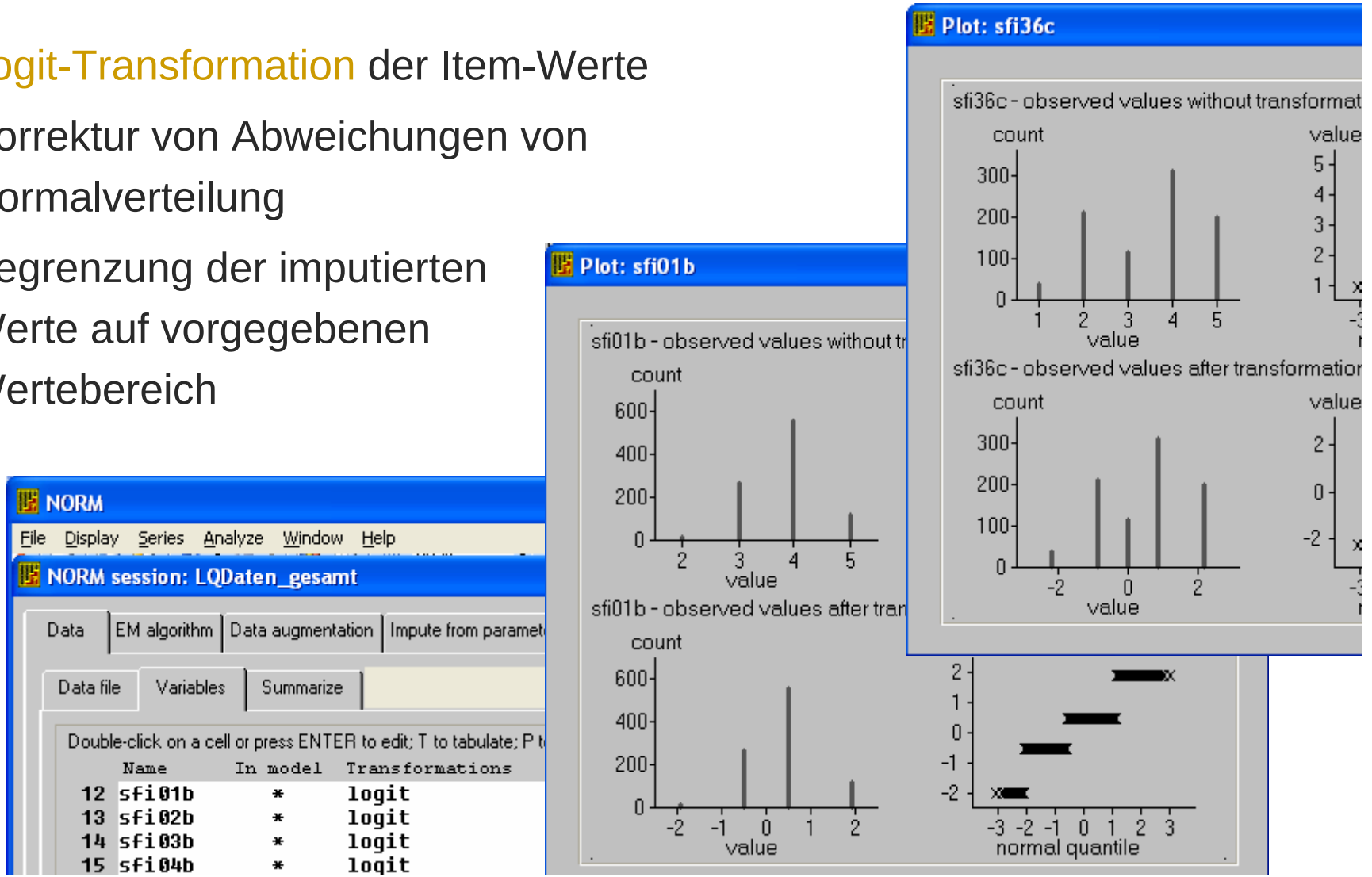
	NUMBER MISSING	% MISSING
dia1	0	0.00
sex	178	15.55
alter	187	16.33
soz1	186	16.24
soz3	425	37.12
soz4	192	16.77
soz5	325	28.38
soz10	227	19.83
gewicht	197	17.21
groesse	200	17.47
sfi01b	188	16.42
sfi02b	184	16.07
sfi03b	187	16.33
sfi04b	189	16.51
sfi05b	187	16.33
sfi06b	191	16.68
sfi07b	195	17.03
sfi08b	189	16.51
sfi09b	192	16.77

Datenaufbereitung

- Logit-Transformation der Item-Werte

⇨ Korrektur von Abweichungen von Normalverteilung

⇨ Begrenzung der imputierten Werte auf vorgegebenen Wertebereich

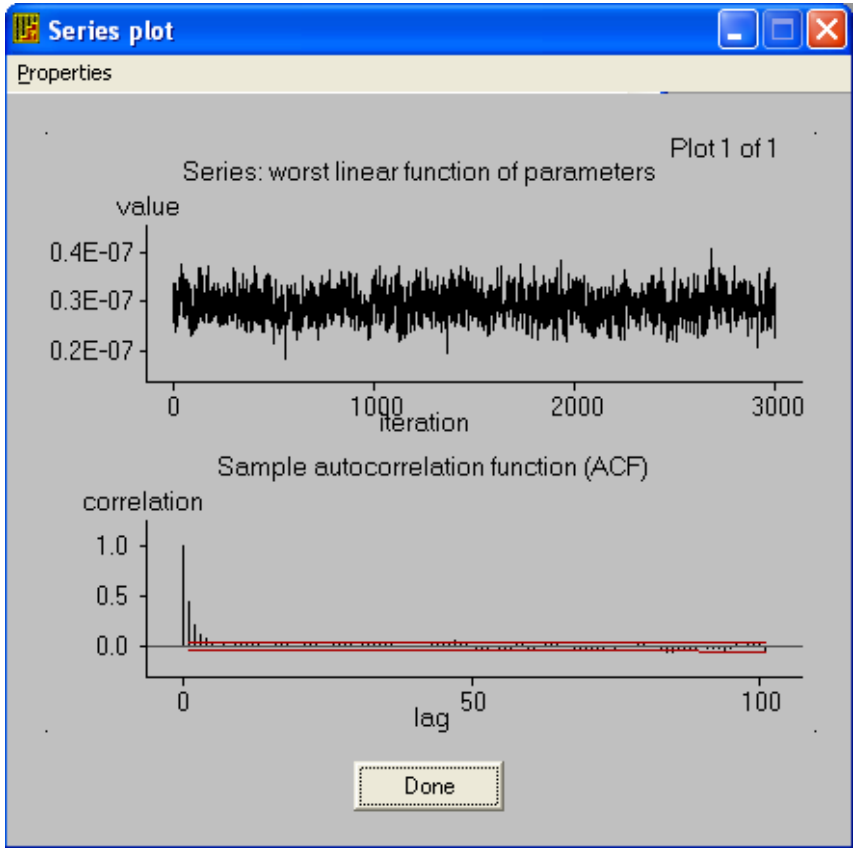


Beurteilung der Konvergenz

- Iterationen: 5 Imputationen x 300 Iterationen = 1500 Iterationen

- Worst Linear Function:
  keine Trends

- AutoCorrelation Function (ACF):
  keine bedeutsame Korrelation
  (p<.05)



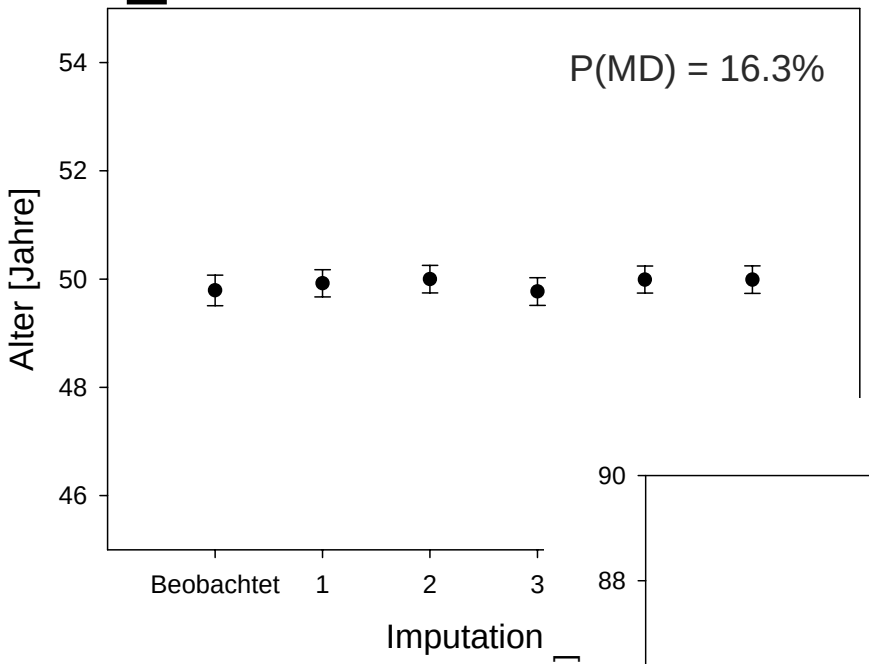
Effizienz

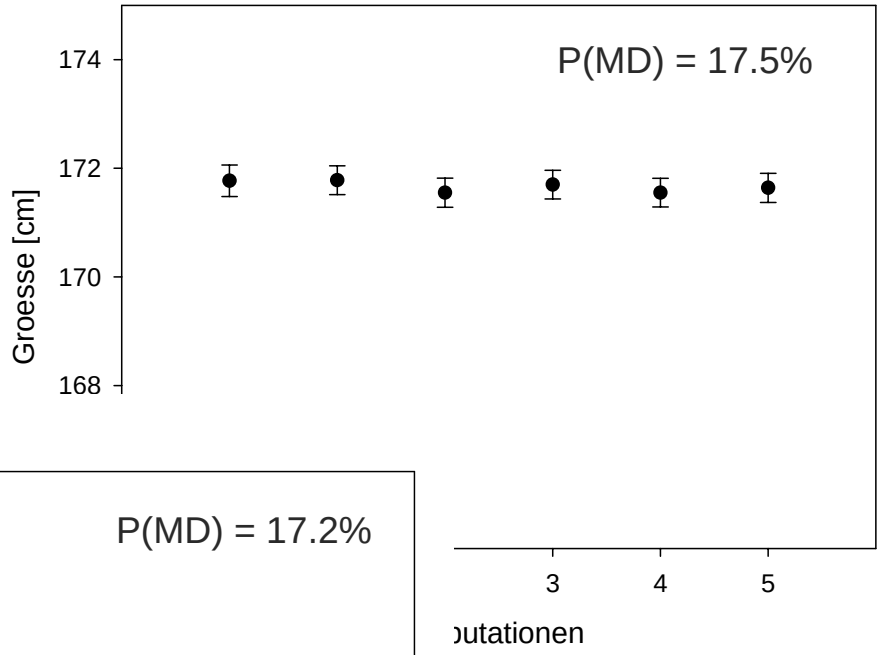
Impu-tationen	Anteil fehlender Information (γ)				
	0.1	**0.3**	0.5	0.7	0.9
3	97	91	86	81	77
5	98	**94**	91	88	85
10	99	97	96	93	92
20	100	99	98	97	96

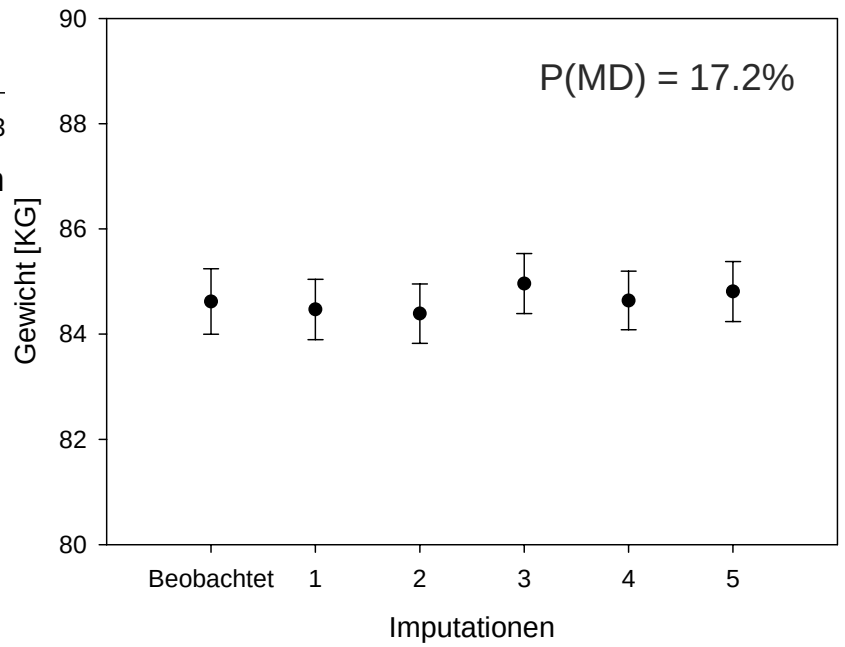
Tab. 1: Effizienz (in %) von MI abhängig von fehlender Information (γ) und Anzahl der Imputationen

III. Ergebnisse

Imputation: Biometrische Daten

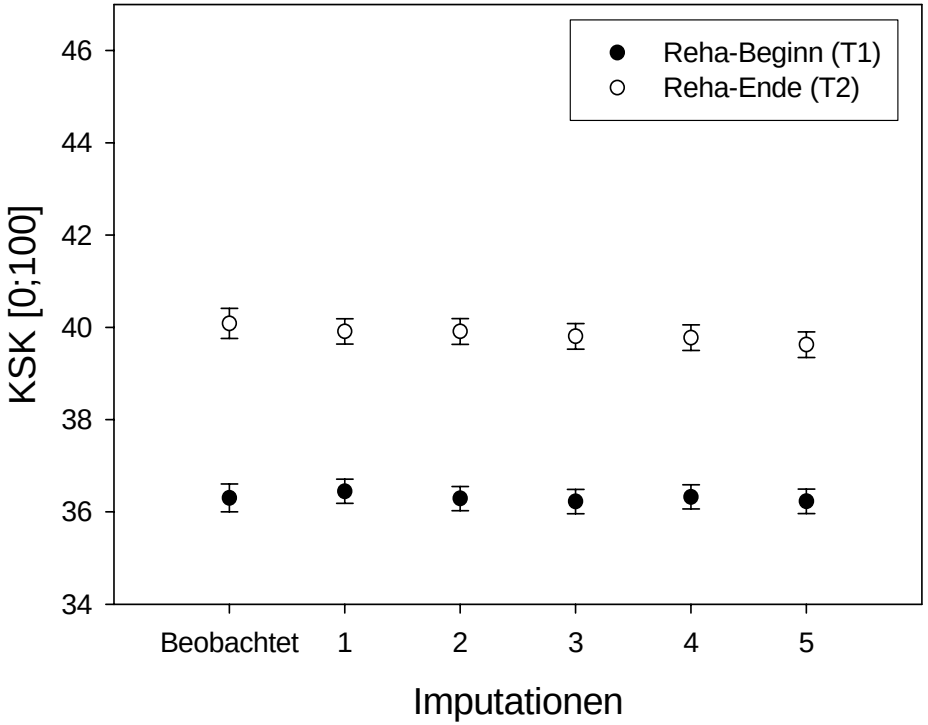






Imputation: SF-36 Summenskalen

Körperliche Summenskala (KSK)

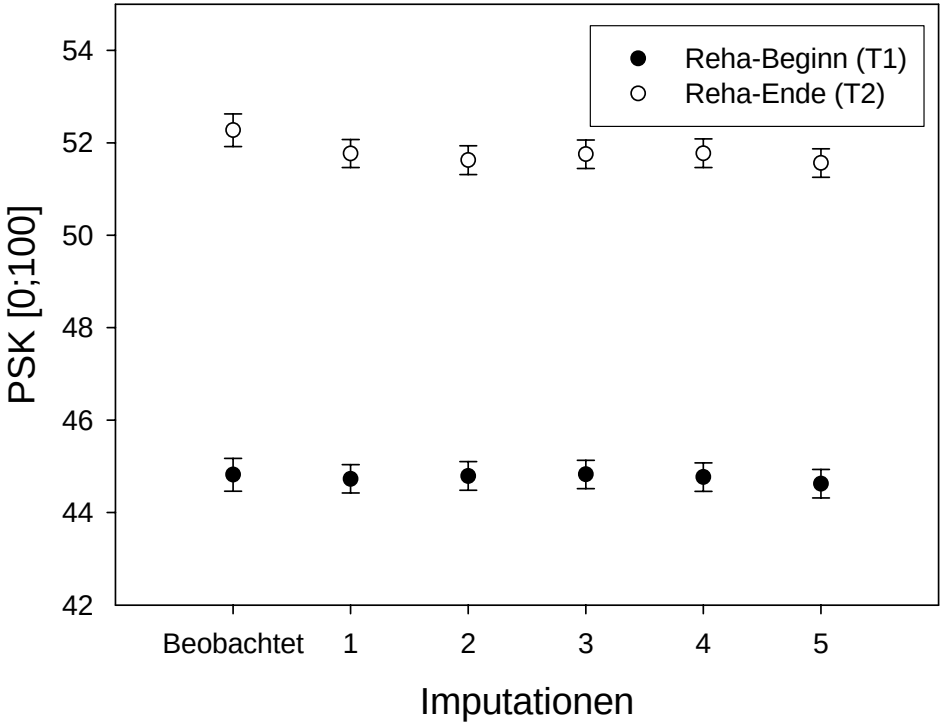


Anteil fehlender Information (γ):

AFI(KSK-T1) = 13.1%

AFI(KSK-T2) = 19.3%

Psychische Summenskala (PSK)



Anteil fehlender Information (γ):

AFI(PSK-T1) = 5.4%

AFI(PSK-T2) = 11.0%

Vergleich Missing-Data-Methoden (I)

SF-36	Listenweiser Fallausschluss				Multiple Imputation				Differenz Konf. Int. (95%) LF - MI
Summen-skalen	N	M	Konfidenz-intervall (95%)		N	M	Konfidenz-intervall (95%)		
KSK-T1	777	**36.51**	35.86	37.16	1145	**36.30**	35.75	36.86	**0.19**
PSK-T1	777	**45.05**	44.21	45.90	1145	**44.75**	44.03	45.47	**0.25**
KSK-T2	777	**40.30**	39.61	40.95	1145	**39.81**	39.20	40.41	**0.13**
PSK-T2	777	**52.31**	51.59	53.04	1145	**51.70**	51.06	52.34	**0.17**

Vergleich Missing-Data-Methoden (II)

SF-36	Paarweiser Fallausschluss				Multiple Imputation				Differenz Konf. Int. (95%) PF - MI
Summen-Skalen	N	M	Konfidenz-intervall (95%)		N	M	Konfidenz-intervall (95%)		
KSK-T1	902	**36.30**	35.71	36.90	1145	**36.30**	35.75	36.86	**0.08**
PSK-T1	902	**44.81**	44.03	45.61	1145	**44.75**	44.03	45.47	**0.14**
KSK-T2	847	**40.09**	39.44	40.73	1145	**39.81**	39.20	40.41	**0.08**
PSK-T2	847	**52.28**	51.58	52.97	1145	**51.70**	51.06	52.34	**0.11**

IV. Zusammenfassung

Zusammenfassung

- Multiple Imputation ist konventionellen MD-Verfahren überlegen
- Praktische Anwendung ist möglich mit SPSS und NORM (Freeware)
- Ergebnisse von Multipler Imputation vs. konventionellen Methoden
  - große Unterschiede der Fallzahlen (777 < 800/950 < 1145)
  - keine signifikanten statistischen Abweichungen
  - keine inhaltlich bedeutsamen Unterschiede
- Aussagen bezogen auf:
  - aggregierte Lebensqualitätsdaten im Verlauf einer Rehabilitation
  - großer Datensatz
  - univariate Auswertungen
  - moderater Anteil fehlender Werte bzw. Information

Ausblick

Weitere Forschung notwendig für:

- **kleine Datensätze (N < 400)**
- **hohen Anteil fehlender Werte (> 30%)**
- **multivariate Auswertungen**

Literatur

1. Igl, W. (2004). Behandlung von fehlenden Werten mit Multipler Imputation. Vortrag beim 13. Rehabilitationswissenschaftlichen Kolloquium „Selbstkompetenz: Weg und Ziel der Rehabilitation“ vom 08. bis 10. März 2004 in Düsseldorf. Quelle: http://www.uni-wuerzburg.de/rehabilitation/methodenberatung/publikationen.html
2. Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. *Rehabilitation, 43*, 109-115.
3. Schafer, J.L. & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7* (2),147-177
4. Sinharay, S., Stern, H.S. & Russell, D. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods*, 6(4), 317-329.
5. Schafer, J.L. (2000). *NORM 2.03 for Windows 95/98/NT* [Software]. Quelle: http://www.stat.psu.edu/~jls

Vielen Dank

für Ihre Aufmerksamkeit!

Kontakt: wilmar.igl@mail.uni-wuerzburg.de

Anhang

1) Imputation: DA-Algorithmus (II)

- I-Schritt: $Y_{mis}^{t+1} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)})$
- P-Schritt: $\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$
- Markov-Kette: $Y_{mis}^{(1)}, \theta^{(1)}, Y_{mis}^{(2)}, \theta^{(2)}, \ldots, Y_{mis}^{(t)}, \theta^{(t)}$
- Endverteilung: $P(Y_{mis}, \theta \mid Y_{obs})$

3) Integration (nach Rubin, 1987)

- MI Punktschätzer: $\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$
- Varianz(innerhalb): $\overline{U} = \frac{1}{m}\sum_{i=1}^{m}U_i$
- Varianz(zwischen): $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2$
- Varianz(gesamt): $T = \overline{U} + (1+m^{-1})B$

3) Integration (II)

- MI Konfidenzintervall: $\overline{Q} \pm t_{df} \sqrt{T}$
- Freiheitsgrade (df): $df = (m-1)(1+\frac{m\overline{U}}{(m+1)B})^2$
- Anteil fehlender Information: $\gamma = \frac{r + 2/(df+3)}{r+1}$

$$r = (T - \overline{U})/\overline{U}$$