OPEN ACCESS

Originalarbeit

# Software zur Behandlung und Ersetzung fehlender Werte

## Software for handling and replacement of missing data

## Abstract

In medical research missing values often arise in the course of a data analysis. This fact constitutes a problem for different reasons, so e.g. standard methods for analyzing data lead to biased estimates and a loss of statistical power due to missing values, since those methods require complete data sets and therefore omit incomplete cases for the analyses. Furthermore missing values imply a certain loss of information for what reason the validity of results of a study with missing values has to be rated less than in a case where all data had been available. For years there are methods for replacement of missing values (Rubin, Schafer) to tackle these problems and solve them in parts. Hence in this article we want to present the existing software to handle and replace missing values on the one hand and give an outline about the available options to get information on the other hand. The methodological aspects of the replacement strategies are delineated just briefly in this article.

**Keywords:** missing values, missing value software

**Benjamin Mayer**[1]
**Rainer Muche**[1]
**Kathrin Hohl**[2]

1 Institut für Biometrie, Universität Ulm, Ulm, Deutschland

2 Biberach a.d. Riss, Deutschland

## Zusammenfassung

In der medizinischen Forschung treten im Zuge einer Datenanalyse oftmals fehlende Werte auf. Dieser Umstand stellt aus verschiedenen Gründen ein Problem dar: Aufgrund fehlender Werte führen beispielsweise Standardmethoden für die Analyse von Daten zu verzerrten Schätzern und einem Powerverlust, da diese vollständiges Datenmaterial voraussetzen und deshalb die unvollständigen Fälle nicht berücksichtigen. Zudem bedeuten fehlende Werte einen gewissen Informationsverlust, weshalb die Aussagekraft der Ergebnisse einer Studie mit fehlenden Werten als geringer zu bewerten ist, als wenn alle Daten zur Verfügung gestanden hätten. Seit einigen Jahren gibt es Methoden zur Ersetzung fehlender Werte (Rubin, Schafer), um diese Probleme anzugehen und teilweise zu lösen. Mit diesem Artikel möchten wir daher zum einen die verfügbare Software zur Behandlung und Ersetzung fehlender Werte vorstellen und zum anderen eine Übersicht geben über die vorhandenen Informationsmöglichkeiten. Dabei werden die methodischen Aspekte der Ersetzungsstrategien nur kurz umrissen.

**Schlüsselwörter:** fehlende Werte, Software für fehlende Werte

## 1 Einleitung

In nahezu allen klinischen Studien stellen fehlende Werte einen problematischen Aspekt dar. Das Ziel einer vollständigen Datenerhebung kann nur selten erreicht werden, da aufgrund verschiedenster Ursachen zumindest einzelne Fehlwerte nicht immer vermieden werden können. Gewöhnliche Auswertungsmodelle, wie sie in den meisten statistischen Standardsoftwareprodukten implementiert sind, basieren jedoch auf einem vollständigen Datensatz der erfordert, dass für alle Variablen jeder einzelne Wert vorhanden ist. Im Falle eines fehlenden Wertes muss deshalb die betreffende Beobachtungseinheit, z.B. ein Patient, aus dem Auswertungskollektiv gestrichen werden, wenn die Daten mit Hilfe der Standardpakete analysiert werden sollen.

Dieser Ansatz bezeichnet die Vorgehensweise der so genannten Complete Case Analyse (CCA), bei der nur vollständig erhobene Beobachtungseinheiten für die Datenanalyse berücksichtigt werden. Die CCA bringt jedoch eine ganze Reihe bedeutender Probleme mit sich, die ihre Verwendbarkeit mehr als in Frage stellen: Die Nichtbe-

rücksichtigung ganzer Beobachtungseinheiten kann dazu führen, dass die Fallzahl drastisch reduziert wird, die Variabilität der Merkmale sich verändert, die Aussagekraft der Studie vermindert wird und Parameterschätzer aufgrund der evtl. zerstörten Strukturgleicheit verzerrt sind. Darüber hinaus steht sie in Widerspruch zu einem sehr angesehenen Auswertungsprinzip, der so genannten Intention-to-treat-Analyse, bei der alle Studienteilnehmer entsprechend der Randomisierung auszuwerten und für die Analyse zu berücksichtigen sind. Unter Beachtung der genannten Gründe ist es umso verwunderlicher, dass die CCA dennoch häufig angewandt wird [14], [26].

Beispielhaft für die z.T. enorme Reduktion der Fallzahl bei CCA betrachte man einen Datensatz mit 25 Variablen, wobei (nur) 3% der Werte je Variable zufällig fehlen mögen. Unter der Annahme, dass die fehlenden Werte über den Datensatz hinweg gleichverteilt sind, werden demnach $1-0.97^{25}=0.53$ der Beobachtungen, also mehr als die Hälfte, nicht berücksichtigt. Je größer der Anteil an fehlenden Werten und je größer die Anzahl an Variablen ist, desto größer ist die Wahrscheinlichkeit für einen Powerverlust bei statistischen Verfahren.

Die größte Problematik fehlender Werte ist die mögliche Verzerrung der Ergebnisse und die resultierende Verringerung der Aussagekraft der Studie. Die Verzerrung kann sich auf die geschätzten Behandlungsunterschiede beziehen, die Vergleichbarkeit der Studienarme beeinflussen und die Repräsentativität des Auswertungskollektivs in Frage stellen (so genannter Selektionsbias). Wenn beispielsweise alle Patienten mit einem geringen (keinem) Therapieerfolg in der Placebogruppe die Studie abbrechen und nur diejenigen in der Studie verbleiben, die sich zumindest teilweise verbessern, so kann der tatsächliche große Behandlungsunterschied bei einer CCA nicht festgestellt werden, da die Daten für den Behandlungsmisserfolg in der Auswertung nicht berücksichtigt werden.

Fehlende Werte führen vor allem dann zu nicht vergleichbaren Studienarmen oder zu einem nichtrepräsentativen Auswertungskollektiv (im Vergleich zur Grundgesamtheit), wenn die fehlenden Werte systematisch auftreten. Die Aussagekraft der Ergebnisse ist in derartigen Situationen stark eingeschränkt.

Speziell in der Auswertung großer epidemiologischer Datensätze ist die Durchführung einer CCA sehr problematisch. Die epidemiologischen Auswertungsmodelle enthalten in der Regel eine relativ große Anzahl von Einflussgrößen, um die Strukturgleichheit der primär interessierenden Risikogruppen zu sichern. Je mehr Variablen das Modell jedoch enthält, desto größer ist die Wahrscheinlichkeit, dass bei einer der Variablen ein fehlender Wert auftritt und somit die gesamte Beobachtung in der Auswertung nicht berücksichtigt wird. Mit zunehmender Anzahl an Einflussgrößen reduziert sich daher die Fallzahl entsprechend dem vorab genannten Beispiel, was sich unmittelbar auf die Power auswirkt.

Um das Problem fehlender Werte angemessener behandeln zu können als im Zuge einer CCA, wurden in den letzten Jahrzehnten verschiedene Ersetzungsstrategien zur Behandlung fehlender Werte entwickelt. Dabei wird im Wesentlichen zwischen der so genannten Single Imputation und der Multiple Imputation unterschieden. Bei Single Imputation wird jeder fehlende Wert durch einen plausiblen Wert ersetzt. Dazu stehen deterministische (Mittelwertersetzung, hot deck/cold deck, Regressionsersetzung) und stochastische Methoden (geschätzter Wert wird um zufälligen Korrekturterm erweitert) zur Verfügung. Bei der Multiple Imputation wird ein fehlender Wert durch mehrere plausible Werte ersetzt auf Basis von Verteilungs- oder MCMC-Methoden. Darüber hinaus existieren auch noch modellbasierte Strategien zum Umgang mit fehlenden Werten, welche die Missing Values nicht explizit ersetzen, jedoch den zu Grunde liegenden Mechanismus der fehlenden Werte im Datenmodell berücksichtigen.

Dieser Artikel soll eine Auflistung geeigneter Softwarelösungen mit entsprechenden Anmerkungen für die Bearbeitung von fehlenden Werten in realen Datensätzen sein, die jedoch nicht vollständig sein kann. Die Autoren haben sich nicht zum Ziel gesetzt, zudem einen ausführlichen Ergebnisvergleich der verschiedenen Softwaretools zu präsentieren, der einen deutlich höheren Arbeitsaufwand bedeutet und den Rahmen dieser Veröffentlichung überschritten hätte. Die Software unterliegt einer ständigen Weiterentwicklung (Verbesserung?) mit neuen Versionen und/oder neuen Methoden. Die nachfolgend genannten Programme und zugehörigen Internet-Adressen geben den Stand von Februar 2009 und die Erfahrungen der Autoren mit den verschiedenen Systemen wieder. Aus den angegebenen Quellen zitierte Informationen konnten nicht alle geprüft werden und sollten von Anwendern dementsprechend mit Vorsicht behandelt werden. In wieweit die Programme und Routinen validiert sind, sollte in der jeweiligen Dokumentation stehen und dort nachgelesen werden.

Der Artikel ist wie folgt aufgebaut: Zu Beginn wird in Abschnitt 2 ein Überblick gegeben zur Diagnostik fehlender Werte, außerdem werden die wichtigsten Ersetzungsstrategien der Single Imputation und der Multiple Imputation vorgestellt. Anschließend werden im Abschnitt 3 einige Internetseiten angegeben, die neben vielen Informationen zu fehlenden Daten auch Übersichten über Softwarelösungen präsentieren. Dies geschieht in der Annahme, dass diese Seiten von den jeweiligen Autoren weiter gepflegt werden. Im Abschnitt 4 werden zwei Spezialprogramme zum Umgang mit fehlenden Werten präsentiert: das Programm NORM und die kommerzielle Software SOLAS. Inzwischen werden auch in den großen, bekannten Statistiksoftwarepaketen Lösungen für den Umgang mit fehlenden Werten angeboten. Im Abschnitt 5 werden Lösungen für SAS, SPSS, S-Plus/R und STATA beschrieben. Weitere Softwarelösungen, die für die eine oder andere Anwendungssituation geeigneter sein können, werden abschließend überblicksmäßig im Abschnitt 6 aufgelistet. Am Ende des Artikels findet sich dann eine kurze Zusammenfassung der vorgestellten theoretischen und praktischen Aspekte, sowie einige Empfehlungen zur Nutzung der Software.

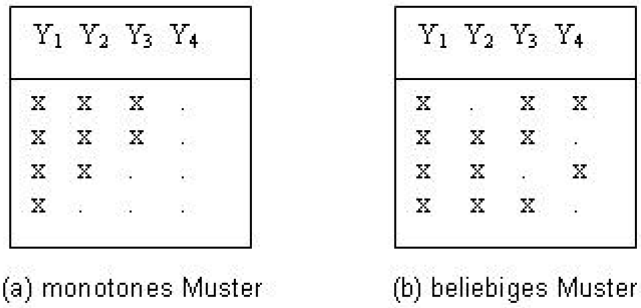


**Abbildung 1: Formen des Missing Data Pattern**

## 2 Missing Data Diagnostic und Ersetzungsstrategien

Die Aussagekraft von Studienergebnissen basierend auf einem Datensatz mit (ursprünglich) fehlenden Werten hängt stark von den Ergebnissen der Missing Data Diagnostic ab. Ein Teil davon besteht aus der Beschreibung, bei welcher Variablen bzw. Beobachtung wie viele fehlende Werte auftreten. Anhand dieser Ergebnisse können mögliche Fehler bei der Dateneingabe oder beim Datenmanagement erkannt werden, die sich eventuell korrigieren lassen.
Zusätzlich werden Unterschiede in der Zielgröße und den charakteristischen Eigenschaften zwischen Beobachtungen mit und ohne fehlende Werte analysiert. Das bedeutet, es wird untersucht, ob fehlende Werte vermehrt bei beispielsweise Alten, Männern oder Rauchern etc. auftreten.
Der andere Teil der Missing Data Diagnostic beschreibt die Anordnung der fehlenden Werte im Datensatz, dem so genannten Missing Data Pattern, und den (möglichen) Gründen für das Auftreten der fehlenden Werte, dem so genannten Missing Data Mechanism. Letzteres ist wichtig für die Wahl einer geeigneten Ersetzungsmethode.
Bei der Bestimmung des Pattern unterscheidet man im Wesentlichen zwischen zwei Mustern. Fehlen die Werte breit gestreut und mehr oder weniger vereinzelt über den ganzen Datensatz hinweg, so spricht man von einem beliebigen oder auch nicht-monotonen Muster. Im Gegensatz dazu steht ein monotones Muster, bei dem die Daten so angeordnet werden können, dass bis zum Beobachtungsende alle Werte eines Merkmals ab einem bestimmten Zeitpunkt, zu dem ein Fehlwert das erste Mal aufgetreten ist, fehlen (Abbildung 1).
Die drei verschiedenen Ausprägungen des Missing Data Mechanismus seien hier nur kurz erwähnt, für eine genauere Beschreibung siehe [10] oder auch [16]. Man unterscheidet in drei Kategorien: Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR). Bei MCAR ist die Drop-out-Wahrscheinlichkeit in keinster Weise abhängig von den Werten der Zielgröße. MAR heißt, dass die Drop-out-Wahrscheinlichkeit nur von den beobachteten Werten abhängt, wobei MNAR bedeutet, dass die Wahrscheinlichkeit für Drop-out (auch) von fehlenden Werten abhängt. Allerdings ist es nahezu unmöglich, den vorliegenden Mechanismus explizit zu identifizieren und in den realen Daten nachzuweisen. Oftmals kann keine strikte Abgrenzung eines bestimmten Mechanismus vorgenommen werden, da es sich um eine Mischform handelt. Zusammen mit dem Pattern bildet dann der Mechanismus den so genannten Missing Data Prozess.
Um mit Standardverfahren der statistischen Datenanalyse arbeiten zu können, bedarf es also im Falle eines unvollständigen Datensatzes einer Ersetzung der fehlenden Werte, wenn man auf eine CCA verzichten möchte. Dafür bieten sich so genannte Single oder Multiple Imputationsverfahren an.
Bei der Single Imputation (SI) wird jeder fehlende Wert durch einen plausiblen Wert ersetzt und daher nur ein vervollständigter Datensatz erzeugt. Zum Beispiel führen alle deterministischen Ersetzungsmethoden eine Single Imputation durch. Das sind Methoden, bei denen die Ersetzung eines fehlenden Wertes durch eine einfache, eindeutige Zuordnung erfolgt. Denkbar sind in diesem Zusammenhang Ersetzungen auf Basis des Mittelwertes bzw. des Medians der beobachteten Daten. Auch so genannte Hot-Deck und Cold-Deck-Techniken kommen ebenso zum Einsatz wie Regressionsverfahren oder stochastische Ersetzungsmethoden [10], [14].
Bei der Multiple Imputation (MI) wird ein fehlender Wert durch mehrere ($m$>1) plausible Werte ersetzt, sodass m vervollständigte Datensätze aus der Ersetzung resultieren. Diese Datensätze werden einzeln mit der gleichen Auswertungsmethode und einem üblichen Softwareprogramm basierend auf einem komplett erhobenen Datensatz ausgewertet. Anschließend werden die Ergebnisse dieser Analysen zu gemeinsamen Schätzern und Standardfehlern zusammengefasst. Das Vorgehen der MI ist in Abbildung 2 graphisch dargestellt und in Little & Rubin [14] genauer erläutert.
Der entscheidende Vorteil der MI gegenüber den SI-Verfahren ist die korrekte Berücksichtigung der Standardfehler. Allen SI-Verfahren ist gemein, dass sie von einer zu geringen Varianz ausgehen. Dem entgegen steht die MI, welche die eigentliche Ersetzung als zusätzliche Varianzquelle richtigerweise beachtet. Demzufolge werden auch Konfidenzintervalle und p-Werte korrekter berechnet, als das mit einer beliebigen SI-Methode möglich wäre.

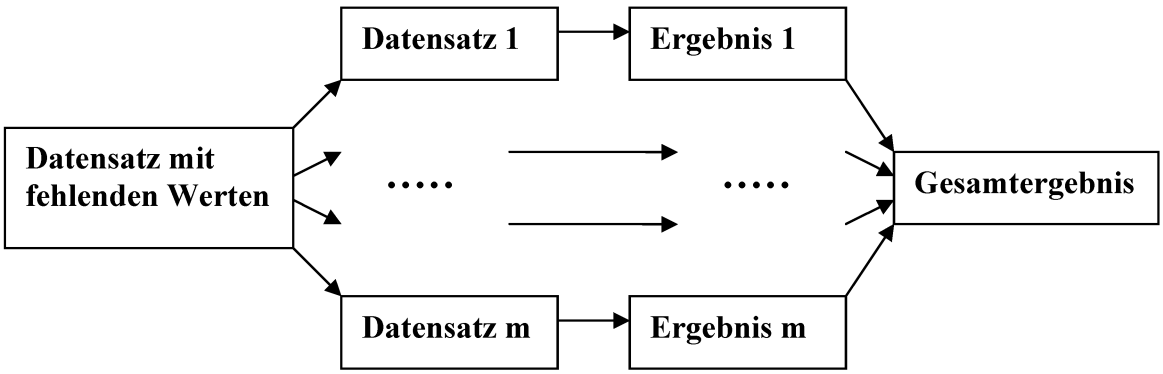


**Abbildung 2: Schema der Multiple Imputation**

# 3 Informationsmöglichkeiten über Softwarelösungen

Informationsmöglichkeiten über geeignete Softwarelösungen lassen sich hauptsächlich in methodischen Fachzeitschriften und im Internet finden.

Artikel in Fachzeitschriften zeichnen sich dadurch aus, dass sie meist unter einem speziellen Aspekt statistischer Analysen die Behandlung fehlender Werte beschreiben und kommentieren. So kann das allgemeine Vorgehen für die eigenen Analysen adaptiert werden, die Softwarehinweise sollten aber möglichst auf Aktualität überprüft werden. Als Beispiel dient hier der Artikel von Horton und Lipsitz [11] aus *The American Statistician* mit dem Titel: „Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables". Dieser Artikel beschränkt sich auf die Beschreibung und den Vergleich von Software für Multiple Imputation ausschließlich im Zusammenhang mit Regressionsmethoden. Ersetzungssoftware ohne statistische Analysemethoden – wie z.B. NORM oder SOLAS – werden bei diesem Vergleich außen vor gelassen. Allerdings geben die Autoren Hinweise auf solche Pakete, so dass der Artikel trotzdem als Startpunkt für eigene Recherchen geeignet wäre. Ein weiteres Beispiel ist der Artikel von Hox [12] mit dem Titel „A Review of Current Software for Handling Missing Data", in dem die Programme SPSS, SOLAS und NORM benutzt werden.

Die Informationsmöglichkeiten über Softwarelösungen zur Behandlung fehlender Werte sind in der Regel aber aktueller im Internet als z.B. in Artikeln in Fachzeitschriften, die schnell veralten. Da Anwendungssoftware oft und schnell verändert wird, ist eine Aktualität in den Informationen dringend notwendig.

Bei den Recherchen zu diesem Artikel wurden wir auf verschiedene Seiten im Internet aufmerksam, die als Startpunkt für einen eigenen Überblick zum Thema Missing Data dienen können. Die Seite http://www.missingdata.org.uk von James Carpenter und Mike Kenward (London School of Hygiene & Tropical Medicine) bietet einen umfangreichen Überblick zu fehlenden Werten mit vielen Beispielen und Grundlagen. Unter der Adresse http://methodology.psu.edu stellt das Team um Linda Collins und Joseph Schafer von der Pennsylvania State University ebenfalls einen Überblick des Themas zur Verfügung. Die Abteilung von Sheldon Ekland-Oslon, University of Texas, macht unter der Internetseite http://ssc.utexas.edu/consulting/answers/general/gen25.html allgemeine Angaben zu speziellen Fragen, die im Bereich fehlende Werte auftreten. Zu guter Letzt erreicht man unter http://www.multiple-imputation.com die Seite von Stef van Buuren, Leiter des TNO in Leiden, Dep. of Statistics, mit umfangreichen Informationen zu Multiple Imputation und entsprechender Software.

Darüber hinaus kann nur der Hinweis auf Suchmaschinen für wissenschaftliche Artikel der betreffenden Methodik in Fachzeitschriften gegeben werden.

# 4 Spezialsoftware für den Umgang mit fehlenden Werten

Es gibt mehrere spezielle Softwarelösungen für den Umgang mit fehlenden Werten, insbesondere der Multiple Imputation. Diese Softwarepakete führen hauptsächlich die Ersetzung der fehlenden Werte aus und geben die Datensätze zur weiteren Auswertung aus. Die eigentliche Analyse kann dann mit der vom Analysten gewohnten Statistiksoftware durchgeführt werden. Neben der Datenübergabe (Eingabe und Ausgabe) an das Ersetzungsprogramm hat man dann aber selber nach der Analyse der m ersetzten Datensätze im Zuge einer Multiple Imputation für die geeignete Zusammenfassung der m Einzelergebnisse zu sorgen, speziell die Zusammenfassung der Varianzkomponenten ist von Bedeutung [14]. In diesem Kapitel wird zunächst der Einsatz der von Schafer entwickelten, kostenlosen Public-Domain Software NORM beschrieben. Im darauf folgenden Abschnitt 4.2 werden Hinweise auf die kommerzielle Software SOLAS gegeben, die ebenfalls speziell zur Diagnostik und Ersetzung fehlender Werte konzipiert ist. Im Artikel von Deal [5] werden beide Produkte verglichen und Deal kommt zu dem Schluss, das seine „limited investigation has not identified a clear winner between SOLAS 3.2 and NORM 2.03 in terms of satisfying Schafer's goals". Deshalb finden wir es legitim, hier eine Freeware neben einem kommerziellen Produkt darzustellen.

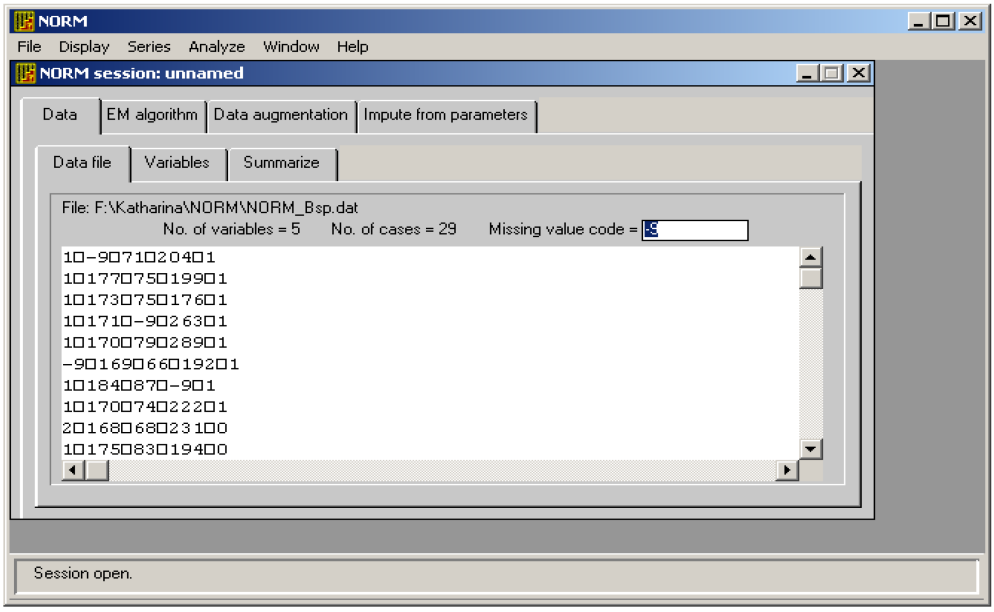


**Abbildung 3: Benutzeroberfläche des Programms NORM**

## 4.1 NORM

NORM wurde von J.L. Schafer entwickelt und basiert auf den multivariat normalverteilten Modellen, die er in seinem Buch [20] beschreibt. Mit NORM kann eine Multiple Imputation durchgeführt werden. Als Ersetzungsmethoden stehen dem Anwender der EM Algorithmus und/oder der Data Augmentation Algorithmus (MCMC) zur Verfügung. Schnelle ad hoc Ersetzungsmethoden (SI) sind nicht implementiert, aber diese sind ja ohnehin nur äußerst selten geeignete Ersetzungsmethoden.

Ein wesentlicher Vorteil von NORM ist, dass es kostenlos von der Internetseite http://www.stat.psu.edu/~jls/misoftwa.html herunter geladen werden kann. Es ist sehr leicht zu erlernen und benutzerfreundlich. Jedoch muss zunächst eine relativ aufwendige Datenaufbereitung vorgenommen werden, um die Daten in NORM einlesen zu können.

Die aktuelle Version NORM 2.03 ist eine eigenständig laufende Software für das Betriebssystem Windows. Schafer schreibt auf seiner Internetseite, dass die Vorgängerversion 2.02 unter Windows 95/98/NT lauffähig ist. Eine Testinstallation von NORM 2.03 unter Windows XP verlief problemlos.

Zur Validierung schreibt Schafer auf der oben angegebenen Internetseite: „This software was written by Joe Schafer (http://www.stat.psu.edu/~jls/) of the Department of Statistics, The Pennsylvania State University. Maren Olsen (same affiliation) assisted in the development of the stand-alone Windows applications. The software may be distributed free of charge and used by anyone if credit is given. It has been tested fairly well, but it comes with no guarantees and the authors assume no liability for its use or misuse."

Die zu bearbeitende Datei muss in NORM importiert werden, da das Programm sich nicht in eine andere Software einbinden lässt. Hierzu wird im Editor eine ASCII-Datei erstellt, die alle Variablenwerte enthält. Jeder Datensatz muss dabei in einer Zeile stehen und die Variablenwerte müssen jeweils durch Leerzeichen oder Tabs getrennt sein. Fehlende Werte sind mit einer geeigneten Codierung numerischen Typs zu versehen. Diese Datei darf keine Variablennamen enthalten und muss als dat-Datei abgespeichert werden. Für die Zuweisung der Werte zu ihren entsprechenden Variablennamen muss eine gleichnamige Datei im nam-Format angelegt werden. Die Variablennamen müssen Zeile für Zeile untereinander geschrieben werden, dann ist diese Datei im selben Ordner zu speichern (Abbildung 3).

Die Ergebnisse von NORM sind vervollständigte Dateien ohne Variablennamen im ASCII-Format. Diese Dateien müssen dann in ein anderes Statistik-Programm exportiert werden, um die Datenanalyse vorzunehmen.

Für weitere Informationen zur Anwendung von NORM kann an dieser Stelle auf eine kurze Anleitung verwiesen werden, die an unserem Institut angefertigt wurde. Bei Interesse setzen Sie sich bitte mit den Autoren in Verbindung.

## 4.2 SOLAS

SOLAS (Version 3.2) ist ein kommerzielles Softwareprodukt zur Behandlung und Ersetzung fehlender Werte. SOLAS wird vertrieben von der Firma Statsol, die auch viele weitere Statistiksoftware im Angebot hat (http://www.statsol.ie/). Das Programm wurde in Zusammenarbeit mit Rubin, dem „Erfinder" der Multiple Imputation, entwickelt und hat sicher den größten Leistungsumfang, wenn es um Diagnostik und Ersetzung fehlender Werte geht.

Neben einigen Varianten der Multiple Imputation hat SOLAS 3.2 auch einige Single Imputationsmethoden bis hin zu deterministischen Ersetzungsmethoden implementiert. Allerdings wird zur Nutzung dieser Methoden auf

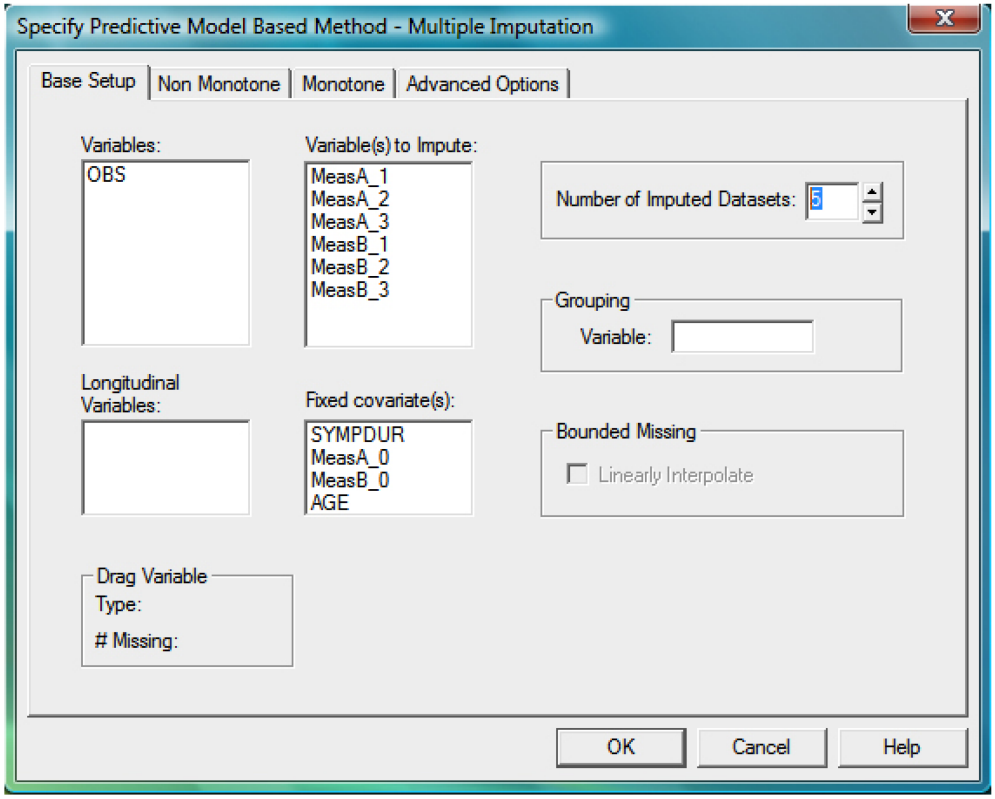


**Abbildung 4: Multiple Imputation mit SOLAS**

der Internetseite mit den wichtigsten Informationen zu SOLAS (http://www.statsol.ie/index.php?pageID=5) Rubin zitiert mit: „*SOLAS is currently the only program that implements multiple imputation noniteratively and with substantial flexibility, even including ad-hoc methods, such as LOCF, as points of comparison for sensitivity analysis.*", so dass er diese Methoden hauptsächlich als geeignet für Vergleichszwecke ansieht (Abbildung 4).
Von der zitierten Internetseite kann eine lauffähige Demo-Version geladen werden und es findet sich umfangreiches Informationsmaterial zur Methodik, Features, Statistik und Datenmanagement. Nach Angaben der Hersteller ist der Datentransport in SOLAS wesentlich einfacher als in NORM, da ein Import/Export von SAS, SPSS, S-Plus, SYSTAT, Stata, BMDP, Excel und ASCII angegeben wird. Der Hauptnachteil der Anwendung von SOLAS wird der Preis sein. Eine Lizenz im akademischen Bereich kostet 995 €, eine Lizenz für kommerzielle Anwendungen 1295 € (Stand Februar 2009). Nach Angaben der Hersteller ist SOLAS validiert. Auf der Internetseite sind einige Dokumente zur Validierung dokumentiert. SOLAS läuft unter Windows ab Version 95.

# 5 Standard-Statistiksoftware und der Umgang mit fehlenden Werten

Die zur Auswertung anstehenden Daten werden üblicherweise mit einem der bekannten und großen Statistiksoftwarepakete analysiert. Davon gibt es sehr viele und es war nicht möglich, alle in Bezug auf Möglichkeiten der Behandlung und Ersetzung fehlender Werte zu prüfen. Daher haben wir hier für die unserer Meinung nach am häufigsten eingesetzten Programmpakete (SAS, SPSS, S-Plus/R, Stata) die entsprechenden Features zusammengestellt. Allerdings haben wir bei unseren zahlreichen Recherchen auch keine wesentlichen Hinweise auf die Behandlung fehlender Werte in anderen Paketen gefunden.

## 5.1 SAS

SAS (Version 9.2) (http://www.sas.com/offices/europe/germany/index.html) ist einer der Marktführer unter den Statistiksoftwarepaketen und wird häufig im Umfeld klinischer Forschung an Universitäten und der pharmazeutischen Industrie eingesetzt. Um die volle Leistungsfähigkeit auszuschöpfen (z.B. in Bezug auf die Ersetzung fehlender Werte) muss die umfangreiche SAS-Syntaxsprache genutzt werden. In den mitgelieferten maus- und menügesteuerten Oberflächen sind die Ersetzungsmethoden nicht abrufbar. SAS bietet die Möglichkeit, in so genannten Makros Unterprogramme in SAS-Syntax zu programmieren, die dann spezielle Auswertungsroutinen zusätzlich zur Verfügung stellen. Dies wird in Kreisen von SAS-Anwendern oft genutzt, so dass neben den offiziellen Prozeduren zur Bearbeitung fehlender Werte auch viele dieser Makros veröffentlicht und verfügbar sind. Die wichtigsten werden nach den Informationen zu SAS eigenen Lösungen hier dokumentiert.

| Methoden | Merkmalstyp | | | Missing Pattern | | Missing Data Mechanism | | durchführbar als | |
|---|---|---|---|---|---|---|---|---|---|
| | nominal | ordinal | stetig | monoton | beliebig | MCAR | MAR | SI | MI |
| Mittelwert/Medianersetzung | | | X | X | X | X | | X | |
| Predictive Mean Matching | | | X | X | | | X | X | |
| Regressionsersetzung | | | X | X | | X | X | X | X |
| EM Algorithmus | | | X | X | X | X | X | X | X |
| MCMC (Data Augmentation) | | | X | X | X | X | X | X | X |
| Logistische Regression | | X | | X | | X | X | X | X |
| Discriminant Function Method | X | | | X | | X | X | X | X |

**Abbildung 5: Ersetzungsmöglichkeiten in %MISSING (Version 9) [8]**

## PROC MI & PROC MIANALYZE

Seit der Version 8.2 hat SAS eine Prozedur zur Durchführung einer Multiple Imputation experimentell eingeführt. Mit dieser Prozedur PROC MI (aktuell in 9.2) lässt sich mittlerweile das Missing Data Pattern ausgeben und die fehlenden Werte mit den Methoden EM-Algorithmus, MCMC-Algorithmus (Data Augmentation), multiple Regressionsersetzung, Logistic Regression Method, Predictive Mean Matching und Discriminant Function Method ersetzen. Dabei wird auf die Vorarbeiten von Allison [1], [2], [3] sowie auf Rubin [19] und Schafer [20] zurückgegriffen. Die Methoden Logistic Regression Method und Discriminant Function Method eignen sich speziell zur Ersetzung von fehlenden Werten kategorialer Variablen [10]. Die Beschreibung der Prozedur kann in der Online-Dokumentation von SAS unter der Adresse http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/mi_toc.htm nachgelesen werden. Ein Auszug aus der Prozedurbeschreibung erklärt:

*The MI procedure performs multiple imputation of missing data...Multiple imputation does not attempt to estimate each missing value through simulated values. Instead, it draws a random sample of the missing values from its distribution. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, confidence intervals with the correct probability coverage.*

Die Veröffentlichung von Yuan [27] beschreibt die Möglichkeiten von PROC MI (http://www.sas.com/rnd/app/papers/multipleimputation.pdf), allerdings noch in der Version 8.2.

Zusätzlich zu PROC MI bietet SAS die Prozedur PROC MIANALYZE an, die für einige Regressionsanalysen (hauptsächlich lineare Regression) die Ergebnisse einer mit PROC MI durchgeführten Multiple Imputation geeignet zusammenführt. Einfache Ersetzungsmethoden wie z.B. die Mittelwertersetzung konnten in einigen weiteren Prozeduren in SAS schon lange durchgeführt werden. Ersetzungsmethoden finden sich u.a. in den Prozeduren PROC STANDARD / PROC STDIZE (Base SAS), PROC PRINQUAL (SAS/STAT) und PROC EXPAND (SAS/ETS).

## SAS Makros zur Bearbeitung fehlender Werte

Im Folgenden werden neben den eigenen Entwicklungen der Autoren die bekanntesten und am häufigsten zitierten SAS-Makros zur Bearbeitung fehlender Werte aufgelistet und kurz beschrieben.

### SAS-Makros von Hohl und Muche: %MISSDESCRIPTION und %MISSING

Hohl und Muche stellen zwei Makros zur Diagnostik und Ersetzung fehlender Werte auf Basis der Prozedur PROC MI zur Verfügung: %MISSDESCRIPTION und %MISSING [4], [9]. Die Makros sind für die SAS-Version 9 von der Internetseite http://www.uni-ulm.de/med/med-biometrie/forschung/sas-makros-fuer-missing-values.html zu beziehen.

Das Makro %MISSDESCRIPTION dient zur Beschreibung eines vorliegenden Datensatzes speziell in Bezug auf fehlende Werte. Zunächst wird der Anteil an fehlenden Werten je Variable und im gesamten Datensatz, optional die Beobachtungen mit den meisten fehlenden Werten und anschließend das Missing Data Pattern (aus PROC MI) angegeben. Darüber hinaus erfolgt eine (gewöhnliche) Deskription aller Variablen [4], [9].

Mit dem Makro %MISSING können Single und Multiple Imputation durchgeführt werden. Eine Single Imputation bei stetigen Variablen wird unter Nutzung der SAS-Prozedur PROC STDIZE durchgeführt. Fehlende Werte können hierbei durch den Median oder Mittelwert der vorhandenen Beobachtungen ersetzt werden. Bei kategorialen Variablen ist zudem die Erzeugung einer eigenen Missing-Kategorie möglich [4], [9].

Der Leistungsumfang des Makros %MISSING in Bezug auf die im jeweiligen Fall sinnvollen Ersetzungsmethoden ist in Abbildung 5 aufgelistet.

### SAS-Makropaket von Müller: Analyse und Ersetzung von Missing Data

Verschiedene SAS-Makros zur Diagnostik und Ersetzung fehlender Werte werden von Müller auf seiner Internetseite zur Verfügung gestellt, erreichbar unter http://www.joergmmueller.de/AuswahlEntwickelterAnwendungssoftware.htm.

- *%INDIKAT (2000) Erstellung einer Missing-Data Indikatormatrix*
- *%MISSING (2000) Analyse der Missing-Data nach Personen und Variablen*
- *%KATPAT (2000) Analyse der bivariaten Verteilung von Missing Data*

- *%IMPUTAT (1999) Multivariaten Datenersetzung*
- *%CHECKIMP(1999) Kontrolle der ersetzten Werte*

### SAS-Makros von Allison

Einer der wichtigsten Autoren zur Methodik und Anwendung zur Bearbeitung fehlender Werte, Allison [1], [2], [3] stellt seit langem SAS-Makros für Multiple Imputation zur Verfügung. Diese Makros stammen aus der Zeit vor PROC MI und waren u.a. Grundlage bei der Entwicklung der Prozedur. Die folgenden Makros sind von seiner Internetseite http://www.ssc.upenn.edu/~allison/ (SAS Macros) abrufbar:

- *MISS (version 1.05) uses the EM algorithm to estimate the parameters of a multivariate normal distribution when data are missing, and optionally generates multiply imputed data sets using the methods of Schafer.*
- *COMBINE (version 1.03) takes estimates based on multiply imputed data sets and combines them into a single set of estimates and associated statistics.*
- *COMBCHI (version 1.0) takes chi-square statistics from multiply imputed data sets and produces a single p-value.*

### SAS-Makro von Gregorich: EM_COVAR

Steve Gregorich stellt unter http://lib.stat.cmu.edu/general/em_covar.sas ein SAS-Programm EM_COVAR zur Verfügung, mit dem durch die Anwendung des EM-Algorithmus eine ML-Kovarianzmatrix und ein zugehöriger Mittelwertvektor geschätzt werden kann.

### SAS-Makro von van Buuren: MISTRESS

MISTRESS ist eine spezielle Methode zur Ersetzung fehlender kategorialer Daten [23]. Das SAS-IML-Makro MISTRESS V. 1.17 steht unter http://www.stefvanbuuren.nl/mistress/index.html zur Verfügung.

### SAS-Makro von Raghunathan et al.: IVEWARE

IVEWARE (Imputation and Variance estimation) ist ein SAS basiertes Softwarepaket (URL http://www.isr.umich.edu/src/smp/ive/). Mit IVEWARE kann eine Multiple Imputation ähnlich wie MICE (siehe 5.3 S-Plus) durchgeführt werden:

1. *Perform single or multiple imputations of missing values using the Sequential Regression Imputation Method* [18].
2. *Perform a variety of descriptive and model based analyses accounting for such complex design features as clustering, stratification and weighting.*
3. *Perform multiple imputation analyses for both descriptive and model-based survey statistics.*

### SAS-Makro von Brown et al.: SIRNORM

Ein weiteres SAS-Makropaket zur Durchführung einer Multiple Imputation ist SIRNORM (http://web.usf.edu/psmg/Sirnorm/sirnorm.html).

- *sirnorm.sas (plain text): General Purpose – Procedure for Multiple Imputations using the Sampling/Importance Resampling Algorithm (SIR).*
- *mult_inf.sas (plain text): General Purpose – Procedure to perform statistical inference on multiple imputations.*
- *m_var_co.sas (plain text): General Purpose – Procedure for means and the variance/covariance matrix of the mean from multiple imputations.*
- *example1 (plain text): Simple example to show how to use multiple imputations procedure.*

### SAS-Makro von Little und Yau: Multiple Imputation in Zeitverläufen

Little und Yau haben 1996 eine Methode zur Ersetzung fehlender Werte in der speziellen Auswertungssituation longitudinaler Daten mit Drop-Outs (ITT-Analyse in klinischen Studien) vorgeschlagen [15] und dokumentieren die entsprechenden SAS-Programme auf der Internetseite http://www.sph.umich.edu/~rlittle/jobs2.htm.

## 5.2 SPSS

SPSS (aktuelle Version 17.0 unter http://www.spss.com/de/) ist ein in den Sozialwissenschaften und in der Psychologie häufig genutztes Statistikpaket. Mit einer maus- und menügesteuerten Oberfläche ist die Bearbeitung von Datensätzen relativ einfach zu erlernen. Im Basispaket der Software sind nur sehr wenige Möglichkeiten zur Behandlung von fehlenden Werten implementiert. Im Wesentlichen werden Methoden in entsprechenden Zusatzmodulen angeboten. Folgende Zusatzmodule sind einsetzbar für die Diagnostik und Ersetzung fehlender Werte:

### SPSS Data Validation

Das Zusatzmodul Data Validation wird seit der SPSS-Version 14 angeboten (http://www.spss.com/data_preparation) und kann u. a. zur Diagnostik fehlender Werte eingesetzt werden:

- *Identifizieren Sie auffällige oder ungültige Fälle, Variablen und Werte, erkennen Sie Muster in fehlenden Daten und fassen Sie Variablen-Verteilungen zusammen. Bestimmen Sie dann die Validität der Daten und entfernen oder korrigieren Sie verdächtige Fälle nach Belieben vor der Analyse*
- *Entdecken Sie multivariate Ausreißer. Sie können diese weiter prüfen und bestimmen, ob sie in die Analyse miteinbezogen* werden sollen.

### SPSS Missing Value Analysis (MVA)

Das Zusatzmodul MVA ist das Hauptprodukt von SPSS zur Diagnostik und Ersetzung fehlender Werte (http://www.spss.com/de/software/statistics/missing-values/). SPSS beschreibt den Funktionsumfang auf der eigenen Seite so:
*Mit SPSS Missing Value Analysis können Sie Ihre Daten hinsichtlich fehlender Werte analysieren und unter bestimmten Voraussetzungen sogar fehlende Werte durch geschätzte Werte ersetzen. Durch eine Analyse Ihrer Daten auf fehlende Werte vor der eigentlichen Auswertung können Sie überprüfen, ob bestimmte Interpretationen zulässig sind...*
Zu früheren Versionen von SPSS gab es Hinweise, dass die Nutzung dieses Moduls nicht fehlerfrei ist bei der Ersetzung fehlender Werte. So schrieb von Hippel [25], dass mit dem eingefügten EM-Algorithmus nur Single Imputation durchgeführt werden kann:
*In Version 12.0, MVA offers four general methods for analyzing data with missing values. Unfortunately, none of these methods is wholly satisfactory when values are missing at random. The first two methods, listwise and pairwise deletion, are well known to be biased. The third method, regression imputation, uses a regression model to impute missing values, but the regression parameters are biased because they are derived using pairwise deletion. The final method, expectation maximization (EM), produces asymptotically unbiased estimates, but EM's implementation in MVA is limited to point estimates (without standard errors) of means, variances and covariances.*
Ähnliche Erfahrungen mit dem SPSS-Modul MVA (11.0) werden auch von Völkner [24] in seiner Diplomarbeit geschildert. Auf Seite 82 fasst er seine Ergebnisse und Erfahrungen mit der SPSS-Routine folgendermaßen zusammen:
*Die Implementierung des EM-Algorithmus in SPSS ist nur sehr mangelhaft, da die Angabe einer Fallbeschriftung zu Veränderungen der Ergebnisse führt und die – aus der vollständigen Datenmatrix berechneten – Standardabweichungen zu gering ausfallen.*
Ab der aktuellen Version 17.0 stellt SPSS allerdings im Rahmen des MVA-Moduls erstmals die Möglichkeit der Multiple Imputation bei kategorialen oder stetigen Variablen zur Verfügung. Es kann zwischen der MCMC-Methode und einer monotonen Methode gewählt werden. Was unter einer monotonen Methode zu verstehen ist, kann anhand der offiziellen Broschüre (http://www.spss.com/de/media/collateral/statistics/missing-values.pdf) leider nicht nachvollzogen werden. Allgemein steht unter der Adresse http://www.spss.com/statistics/missing_values/ in der offiziellen Ankündigung von SPSS:
*In SPSS Missing Values 17.0, a new multiple imputation procedure will help you understand patterns of "missingness" in your dataset and enable you to replace missing values with plausible estimates. It offers a fully automatic imputation mode that chooses the most suitable imputation method based on characteristics of your data, while also allowing you to customize your imputation model.*

### Amos 6.0

Eine Alternative zum Modul MVA in SPSS stellt die Nutzung von AMOS dar (http://www.spss.com/amos/). AMOS ist eigentlich ein Modul zur Nutzung von Strukturgleichungsmodellen in SPSS, bietet aber darin auch Methoden zur Ersetzung fehlender Werte an. Folgende Funktionalität wird ab SPSS 14.0 auf http://www.spss.com/de/amos beschrieben:

- *Erstellen von Datensätzen mit fehlenden Werten oder latenten Variablen. Verwenden Sie die Regressionsimputation zum Erstellen eines einzelnen, vollständigen Datensets. Die stochastische Regressionsimputation oder die Bayessche Imputation können zum Erstellen mehrerer imputierter Datensets verwendet werden.*

Diese Möglichkeiten sind also nicht im Standardlieferumfang von SPSS enthalten und müssen zusätzlich angeschafft werden.

## 5.3 S-PLUS / R

S-Plus (Version 8.0) http://www.insightful.com/products/splus/default.asp ist ein kommerzielles Statistikpaket, welches sich aus der „Statistiksprache" S entwickelt hat. Parallel wurde dieselbe Sprache weiterentwickelt als Softwarepaket R (Version 2.8.1), welches Public-Domain unter http://www.r-project.org/ erhältlich ist. Beide Softwareprodukte zeigen große Ähnlichkeiten in den Möglichkeiten und in von Anwendern geschriebenen Modulen, so dass sie hier gemeinsam behandelt werden. Wichtig ist allerdings dabei zu beachten, welche Version jeweils als Vorraussetzung für die Anwendungen benötigt wird.
S-Plus beinhaltet seit der Version 8 einige Möglichkeiten für Multiple Imputation. Kollegen vom Hartwell Center for Bioinformatics and Biotechnology fassen dies folgendermaßen zusammen (http://www.hartwellcenter.org/bio_rescom/apps_restools/splus6.php):
*When performing real-world data analysis you often encounter missing values. S-PLUS is the only package that lets you account for the effect of missing values using three different multiple imputation models: Gaussian, Logistic, and Conditional Gaussian. When properly accounted for, missing values can provide critical insight in your analysis and help you leverage your data investment.*
Horton beschreibt in einem seiner Paper [11] die Ersetzungsmöglichkeiten mit S-Plus:
*S-Plus features a new missing data library, which extends S-Plus to support model-based missing data models, by use of the EM algorithm (Dempster, Laird and Rubin 1977) and data augmentation (DA) algorithms (Tanner and Wong 1987). DA algorithms can be used to generate multiple imputations. The missing data library provides support for multivariate normal data (impGauss), categorical data (impLoglin) and conditional Gaussian models*

Tabelle 1: Funktionen in HMISC

| Function Name | Purpose |
|---|---|
| aregImpute | Multiple imputation based on additive regression, bootstrapping and predictive mean matching |
| impute | Impute missing data (generic method) |
| naclus | function for examining similarities in patterns of missing values across variables |

*(impCgm) for imputations involving both discrete and continuous variables.*
Vor diesen Neuerungen seit der Version 6.0 in S-Plus gab es schon viele Jahre von Nutzern entwickelte Libraries für S-Plus und/oder R, mit denen die wichtigsten Ersetzungsmethoden schon frühzeitig mit dieser Software durchführbar waren. Die wichtigsten Beitragenden waren Schafer (NORM) und van Buuren (MICE).

## S-PLUS Programme von Schafer: NORM, CAT, MIX und PAN

Schafer hat schon seit 1999 verschiedene Bibliotheken für S-Plus Versionen zur Verfügung gestellt. Neben der Entwicklung der Stand-alone-Version von NORM (siehe 4.1) sind verschiedene Softwareprodukte für S-Plus entstanden. Auf der Seite über seine Missing Value-Software http://www.stat.psu.edu/~jls/misoftwa.html sind 4 verschiedene Bibliotheken für S-Plus ab Version 3.3 vorhanden:
*At present, four different software packages are available for creating multiple imputations in S-PLUS.*

- *NORM: Multiple imputation of multivariate continuous data under a normal model.*
- *CAT: Multiple imputation of multivariate categorical data under loglinear models.*
- *MIX: Multiple imputation of mixed continuous and categorical data under the general location model.*
- *PAN: Multiple imputation of panel data or clustered data under a multivariate linear mixed-effects model.*

## S-PLUS Programm von van Buuren: MICE

Ein weiteres wichtiges Paket zur Durchführung von Multiple Imputation ist das Programm MICE von der Arbeitsgruppe um van Buuren aus Leiden/Holland [22], mit der eine spezielle MI-Methode umgesetzt wird. MICE ist zu erhalten unter der Seite http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm für S-Plus (ab Version 4.5) und R (ab Version 1.8).
*MICE stands for Multivariate Imputation by Chained Equations. We have written a software library for multiple imputation using S-Plus V4.5 and higher for Windows, and R 1.8 and higher for Windows. The library assists in performing the steps required in a full multiple imputation analysis. There is also an implementation in STATA. Specific features of MICE V1.0 include:*

- *columnwise specification of the imputation model;*
- *arbitrary patterns of missing data;*
- *transformations and index variables;*
- *subset selection of predictors;*
- *supports standard* lm *and* glm *complete-data methods;*
- *automated pooling using the Barnard-Rubin adjustment;*
- *callable user-written imputation functions;*
- *online help files.*

Verschiedene Download-Möglichkeiten für Windows- und Unix-Versionen werden auf der oben angegebenen Internetseite angeboten. Ebenfalls ist ein Download eines Posters möglich, auf dem die wesentlichen Spezifikationen von MICE kurz und knapp nachzulesen sind.

## S-PLUS/R-Bibliothek von Harrell: HMISC

Die S-Plus- und R- Bibliothek HMISC von Harrell [7] enthält einige Funktionen zur Ersetzung von fehlenden Werten (Single und Multiple Imputation) in Kombination mit Analyse- und Kombinationsroutinen (http://lib.stat.cmu.edu/S/Harrell/Hmisc.html). Die neueste Version ist lauffähig ab S-Plus Version 6.0 und/oder R-Version 2.0 und größtenteils in den Lieferumfang der Software übernommen worden (z.B. die Funktion *impute*). Einige der Funktionen aus HMISC sind in Tabelle 1 aufgelistet.

## R-Pakete

Zudem gibt es in R (aktuelle Version 2.8.1) eine ganze Reihe weiterer Pakete, die für den Umgang mit fehlenden Werten konzipiert sind. Die im Folgenden aufgeführten Pakete stehen alle auf der Seite http://cran.r-project.org/web/packages zum Download bereit. Wir geben zu jedem Paket eine kurze Beschreibung, sowie den Autor und das Erscheinungsjahr an, um dem Leser einen Eindruck über die Aktualität der Pakete zu vermitteln.

## R-Paket von Lee et al.: arrayImpute

Dieses Paket wurde von Lee, Yoon und Park im Jahre 2007 zur Verfügung gestellt. Es ist für die Imputation fehlender Werte bei Microarray-Daten erstellt worden.

## R-Paket von Gelman et al.: mi

Das von Gelman, Hill, Yajima, Su und Pittau entwickelte Paket kann seit 2009 genutzt werden. Es wurde konzipiert für die Imputation fehlender Werte und zum Modelltesten.

### R-Paket von Lumley: mitools

Implementiert sind Tools zur Multiple Imputation (Lumley (2008)).

### R-Paket von Gramacy: monomvn

Möglich sind Schätzungen für multivariat normalverteilte Daten mit monotonem Muster der fehlenden Werte (Gramacy (2008)).

### R-Paket von Gross: mvnmle

Ein Paket von Gross (2008) zur ML-Schätzung multivariat normalverteilter Daten mit fehlenden Werten.

### R-Paket von Novo: norm

Novo (2002) hat ein Paket zur Analyse von multivariat normalverteilten Daten mit fehlenden Werten erstellt. Als Basis hierfür dienten die Arbeiten von Schafer.

### R-Paket von Tempi et al.: VIM

Mit diesem Paket von Templ und Alfons (2009) ist es möglich, fehlende Werte zu visualisieren und zu ersetzen.

## 5.4 Stata

Stata (Version 10) ist ein umfangreiches Statistikpaket, was eine große Verbreitung in den USA und weltweit im Bereich der epidemiologischen Forschung hat (http://www.stata.com). Im eigentlichen Softwarepaket Stata gibt es den *impute* Befehl. Die folgende Beschreibung wird angegeben:
*impute fills in missing values; depvar* (http://www.stata.com/help.cgi?depvar) *is the variable whose missing values are to be imputed. indepvars* (http://www.stata.com/help.cgi?indepvars) *is the list of variables on which the imputations are to be based, and newvar1* (http://www.stata.com/help.cgi?newvar) *is the new variable that contains the imputations.*
Auf der Seite http://www.sociology.ohio-state.edu/people/ptv/faq/mi_mianalyze.htm wird im Zusammenhang mit der SAS-Prozedur PROC MI folgende Frage und Antwort veröffentlicht:
*Q: Is Stata's impute command just as good?*
*A: No. It doesn't account for random variation, so it will impute the same value every time. Multiple imputation is based on imputing several random values, and accounting for the variation among them.*
Auf Grundlage dieser Hinweise scheint es, dass das Impute-Statement eine deterministische Regressionsersetzung durchführt.
Stata ist ein Softwarepaket, in das von Nutzern geschriebene Module implementiert werden können.

### Stata-Modul von Royston/van Buuren: MICE

Ein wichtiges Modul ist die Einbindung des Softwarepaketes MICE von van Buuren (siehe 5.3) in Stata durch Royston (http://ideas.repec.org/c/boc/bocode/s446602.html). Dort wird anhand der Befehle mvis und micombine eine Multiple Imputation ermöglicht. Eine Präsentation von Royston auf der Stata-Konferenz 2005 bzgl. der Anwendungsmöglichkeiten und Hintergründe findet sich unter http://www.stata.com/ meeting/11uk/royston.ppt.

### Stata-Modul von Mander und Clayton: HOTDECK

Ein weiteres Modul ermöglicht eine Hotdeck-Ersetzung. Das Modul HOTDECK wurde von Mander und Clayton geschrieben und wird auf der folgenden Seite http://ideas.repec.org/c/boc/bocode/s366901.html genauer beschrieben:
*hotdeck replaces missing values for the variable indicated by its argument. It should be used within a multiple imputation sequence since missing values are imputed stochastically rather than deterministically. The nmiss missing values in each stratum of the data described by the 'by' option are replaced by values sampled from the nobs observed values in the same stratum. The approximate Bayesian bootstrap method of Rubin and Scheker is used; first a bootstrap sample of nobs observations is sampled with replacement from the observed values, and the nmiss missing values are sampled at random (again with replacement) from this bootstrap sample. If a file is specified in a using clause, the modified file is written to disk and the existing data in memory are unchanged. Otherwise the data in memory are modified. This is version 1.65 of the software, requiring Stata v9. hotdeck6 may be used in earlier versions of Stata.*

### Stata-Modul von Millar: LISTMISS

Zur Missing value-Diagnostik kann das von Millar zur Verfügung gestellte Stata-Modul LISTMISS genutzt werden (http://ideas.repec.org/c/boc/bocode/s449506.html):
*listmiss is a post-estimation command that reports the number of missing values for each independent variable. For each independent variable a flag is created to indicate when the variable is missing. The dependent variable is regressed on the missing flag for each independent variable. The statistical significance of the slope is reported as an indicator of whether the dependent variable is statistically different where an independent variable is missing. Another test compares the null model to the model with the missing flag and performs a BIC difference test, again as an indication of whether the dependent variable is statistically different when an independent variable is missing. If the model was specified with robust standard errors, then robust standard errors are used to perform the hypothesis test related to the slope for the missing value flag.*

### Stata-Modul von Cox: NMISSING

Etwas weniger umfangreich ist das Modul NMISSING, mit dem die Anzahl fehlender Werte in den Variablen und/oder in den Beobachtungen ausgezählt werden kann (http://ideas.repec.org/c/boc/bocode/s455901.html). Spezielle Stata-Module (so genannte „ados“) können innerhalb von Stata mit dem „webseek“-Kommando gesucht und nachträglich dazuinstalliert werden. Module im Bereich Missing Data sind beispielsweise folgende:

- *"meanscor" (die Meanscore-Methode bei fehlenden Daten in logistischen Regressionen),*
- *"pattern" bzw. "mvpatt" (Missing Data Pattern Analyse),*
- *"pcamv" (eine PCA mit ML-Schätzung der Kovarianzmatrix bei fehlenden Werten mit Möglichkeit der Imputation)*
- *"regmsng" (Imputationen mit einer Regressionsvariante) und*
- *"whotdeck" (eine gewichtete Variante der Hotdeck-Ersetzung)*
- *“misum“ (deskriptive Statistik der Schätzungen auf Basis von multipler Imputation*
- *“miest“ (Kombination von Ergebnissen über mehrere Datensätze und Berechnung von Schätzungen auf Basis von multipler Imputation)*

# 6 Weitere Software zur Behandlung fehlender Werte

Bei der Recherche der Softwareprodukte zur Behandlung fehlender Werte haben wir neben den bisher dargestellten Informationen zu Spezialsoftware (siehe Kap. 4) und Statistiksoftwarepaketen (siehe Kap. 5) weitere Informationen sammeln können. Diese werden hier ohne größere Kommentare mit entsprechender Angabe der Internetseiten (Stand Februar 2009) wiedergegeben, da sie eventuell für den einen oder anderen Leser wertvolle Hinweise oder Anwendungsmöglichkeiten enthalten. Wir Autoren haben keine Erfahrungen mit diesen Produkten und können dementsprechend keine wertenden Urteile abgeben. Die Softwareprodukte werden hier alphabetisch gelistet.

## 6.1 AMELIA

AMELIA II (http://gking.harvard.edu/stats.shtml) ist ein Windows-Programm, in dem die Arbeitsgruppe um Gary King einen alternativen Algorithmus zur Multiple Imputation einsetzen [12]. Das Programm nutzt einen schnellen EM-Algorithmus, der von NORM abgeleitet wurde, und existiert bereits in zweiter Version 1.2-0. Die Windows-Version braucht nur das Betriebssystem Windows und ist stand-alone. Das Programm ist menü-orientiert. Neben dieser Windows-Version gibt es eine Prozedur für das Statistikpaket GAUSS. Für beide Versionen gibt es sowohl eine PDF- und Online-Dokumentation (http://gking.harvard.edu/amelia/amelia1/docs/). Auf einer weiteren Internetseite http://gking.harvard.edu/projects/miss.shtml finden sich darüber hinaus Informationen aus der Arbeitsgruppe zu fehlenden Werten (Methoden, Software u.a.).

## 6.2 EMCOV

EMCOV (Estimation of Means and Covariances) ist eine von John W. Graham und Scott M. Hofer [6] unter DOS und UNIX laufende Software zur Ersetzung fehlender Werte mit dem EM-Algorithmus und basiert auf den Methoden von Schafer [20] (ftp://ftp.cac.psu.edu/pub/people/jwg4/).

## 6.3 LISREL / PRELIS

LISREL (Version 8.8) ist nach den Angaben des Vertreibers SSI (http://www.ssicentral.com/) die führende Statistiksoftware für Strukturgleichungsmodelle. In der letzten Windowsversion der Software gibt es verschiedene Anwendungen (zitiert von der Internetseite):

- *LISREL for structural equation modeling.*
- *PRELIS for data manipulations and basic statistical analyses.*
- *MULTILEV for hierarchical linear and non-linear modeling.*
- *SURVEYGLIM for generalized linear modeling.*
- *CATFIRM for formative inference-based recursive modeling for categorical variables.*
- *CONFIRM for formative inference-based recursive modeling for continuous variables.*

Im Programmteil PRELIS sind neben Datenmanipulationsmöglichkeiten und einigen Regressionsmethoden die Möglichkeit für *„Imputation by Matching“* und *„Multiple Imputation“* vorhanden. Im Internet ist ein User’s Guide für PRELIS mit entsprechenden genaueren Beschreibungen unter der Firmenseite erhältlich.

## 6.4 MPLUS

MPLUS (Version 5.2) ist ein Statistikpaket für latent-Class-Analysen. Die wesentlichen Informationen zu MPLUS gibt ein einführender Text im Internet unter http://www.ats.ucla.edu/stat/mplus/seminars/IntroMplus/default.htm wieder. MPLUS kann in Bezug auf fehlende Werte neben den dort beschriebenen Auswertungssituationen zur Untersuchung der Datenstruktur in Bezug auf fehlende Werte genutzt werden.

## 6.5 PreScreen

PreScreen (Version 2.1) ist ein Statistikpaket, welches von den Autoren eingebunden wurde in die Software MATLAB 5 als Auswertungstool für MATLAB-Nutzer. Die Hauptanwendungsgebiete sind: (http://www.ncl.ac.uk/CPACTsoftware/PreScreen/)

- *Plotting capabilities: time series/trend plots, scatter plots, normal probability plots, histograms, auto and cross-correlation plots, rank correlation matrix plot, parallel coordinates plot (version 2.1 only)*
- *Variable selection tools based on statistical techniques*
- *Missing values detection & treatment tools*
- *Variable transformations: statistical, filtering, mathematical, time-shifting*
- *Outliers: univariate and multivariate outlier detection & treatments*

Die Möglichkeiten der Untersuchung fehlender Werte sind zusammengefasst unter http://www.ncl.ac.uk/CPACTsoftware/PreScreen/MissingData.html. Die Stärken des Programms liegen sicher in der Untersuchung der fehlenden Werte, direkt zu nutzende Ersetzungsmethoden werden nicht angegeben.

## 6.6 ProMISS

ProMISS ist eine Software zur Ersetzung fehlender Werte mit der Hot-Deck-Ersetzung. Informationen sind erhältlich unter: http://www.atlantecsoftware.com/promiss2.asp.

## 6.7 XMISS

Cyrus R. Mehta, der Entwickler von Statistikpaketen wie STATXACT, LOGXACT kündigt eine Software zum Umgang mit fehlenden Werten im Bereich multipler Regressionsmodelle an. In der Ankündigung wird speziell auf Studien im Umfeld von Tumorerkrankungen eingegangen (Software for missing covariate data in cancer trials), die Methoden sind aber direkt übertragbar auf vergleichbare Auswertungssituationen. (http://cancercontrol.cancer.gov/grants/abstract.asp?applid=6839961):
*This is a Phase II SBIR proposal for completing the development of a comprehensive collection of statistical tools embedded in LogXact, in EGRET, in SAS as PROCs and in SPLUS as functions. This set of tools will compute maximum likelihood estimates for generalized linear models (GLMs) and parametric survival models with missing categorical covariates, where the missing covariates are assumed to be missing at random (MAR). In this Phase II effort, we will expand the current version of tools available in prototype software XMISS to handle: (i) missing categorical covariates for binomial response models with logit, probit, or complementary log-log links, (ii) missing categorical covariates for conditional logistic regression for matched case-control data, (iii) missing categorical covariates for Poisson regression models, (iv) missing categorical covariates for normal linear regression models, (v) missing categorical covariates for ordinal response regression models, (vi) missing categorical covariates for exponential, Weibull and log-normal regression models allowing for right censoring in the response variable...In addition, we will investigate methods for speeding up the EM algorithm as well as develop new algorithms for obtaining good starting values for the EM algorithm. Missing covariate data is very common problem with cancer clinical trials. There exists no commercial software to handle missing covariate data by maximum likelihood method for the range of models listed above.*
Einen Hinweis darauf, dass die Entwicklung der Software XMISS zu Ende geführt wurde, findet sich im Abstracts-Band der International Biometric Society, Eastern North American Region, das anlässlich ihrer Tagung im März 2007 erschienen ist. Unter der URL http://www.enar.org/documents/enar_program_2007.pdf gelangt man zu diesem Abstracts-Band und kann auf Seite 328 etwas über die Anwendung von XMISS lesen.

## 6.8 WinMICE

WinMICE von Jacobusse ist eine Windows-Applikation des S-Plus-Programms MICE von van Buuren (siehe 5.3) und kann kostenlos von der folgenden Internet-Adresse http://web.inter.nl.net/users/S.van.Buuren/mi/docs/WinMICEsetup.exe herunter geladen werden.

# 7 Fazit

In diesem Artikel wurde eine Übersicht gegeben über die vorhandenen Softwaremöglichkeiten zur Behandlung fehlender Werte in klinischen Datensätzen. Nach einer anfänglichen Vorstellung der wichtigsten theoretischen Aspekte folgten Hinweise auf verschiedene Internetadressen, unter denen man allgemeine Information zu Missing Data Software und deren Anwendung erhalten kann.
Der Hauptteil des Artikels bestand allerdings aus einer umfangreichen Zusammenstellung verschiedener Software-Tools, die zur Ersetzung fehlender Werte verwendet werden können. Aufgrund der Vielzahl von Programmen, die dabei präsentiert wurden, wollen wir in dieser Zusammenfassung nochmals die bedeutendsten Programme mit ihren jeweiligen Funktionalitäten in Tabelle 2 darstellen. Es wird darin angegeben, welche Ersetzungsmethoden die betreffenden Programme zur Verfügung stellen und mit welchem Modul dies im Speziellen möglich ist.
Den Autoren war beim Erstellen dieser Tabelle wichtig, dass der Leser sich einen schnellen Überblick verschaffen kann über Softwareprogramme und -komponenten, die für die Behandlung fehlender Werte in Frage kommen. Es ist uns allerdings nicht möglich, explizit auf bestimmte Programme zu verweisen und diese in besonderem Maße zu empfehlen. Die Entscheidung, mit welchem Programm das Problem fehlender Werte bearbeitet wird, hängt zum einen von der Problemstellung und zum anderen vor allem vom Vorwissen des jeweiligen Analysten ab. Nach Ansicht der Autoren sollten Analysten mit Vorkenntnissen in einem der häufig gebräuchlichen Standardsoftware-Pakete wie SAS, SPSS oder R/S-PLUS auch die speziellen Pakete und Module der ihnen vertrauten Software zur Behandlung fehlender Werte nutzen. Untersucher ohne Kenntnis in einer bestimmten Software sind sicherlich gut beraten, wenn sie zur Ersetzung der Missing Values das Open Source-Programm NORM von Joseph Schafer nutzen und

**Tabelle 2: Missing Data Software und ihre Möglichkeiten (MDD=Missing Data Diagnostic, SI=Single Imputation, MI=Multiple Imputation, det. Ersetzung=deterministische Methoden verfügbar)**

| Software | Modul | | MDD | SI | MI | det. Ersetzung |
|---|---|---|---|---|---|---|
| NORM (2.3) | | | | x | x | |
| SOLAS (3.2) | | | | x | x | x |
| SAS (9.2) | PROC MI | | x | x | x | x |
| | %MISSDESCRIPTION | | x | | | |
| | %MISSING | | | x | x | x |
| | von Müller | | x | x | | |
| | von Allison | | | | x | |
| Makros | EM_COVAR | | | x | | |
| | MISTRESS | | | x | | |
| | IVEWARE | | | x | x | |
| | SIRNORM | | | | x | |
| | von Little & Yau | | | | x | |
| SPSS (17.0) | | | x | | x | |
| | Data Validation | | x | | | |
| Module | MVA | | x | | x | |
| | Amos | | | x | x | |
| S-PLUS (8.0) / R (2.8.1) | | | | | x | |
| | CAT | | | | x | |
| | MIX | | | | x | |
| | PAN | | | | x | |
| Pakete | MICE | | | | x | |
| | HMISC | | | x | x | |
| | arrayImpute | | | x | | |
| | mi | | | | x | |
| | mitools | | | | x | |
| | VIM | | | | x | |
| Stata (10.0) | | | | x | | x |
| | MICE | | | | x | |
| Module | HOTDECK | | | x | | |
| | LISTMISS | | x | | | |
| AMELIA | | | | | x | |
| EMCOV | | | | | x | |
| MPLUS | | | x | | | |
| PreScreen | | | x | | | |
| ProMiss | | | | x | | |
| XMISS | | | | | x | |

die vervollständigten Daten dann in mit Hilfe einer anderen bekannten Statistiksoftware weiter verarbeiten. NORM ist relativ benutzerfreundlich konzipiert und die Menge an einstellbaren Programmfeatures wurde überschaubar gehalten, so dass die ersten Ergebnisse schnell erzielt werden können.

Das Forschungsgebiet der Ersetzung von fehlenden Werten ist ein Bereich, der noch viel Entwicklungspotential birgt. Die bestehenden Methoden werden ständig verbessert und bieten Ansatzpunkte zur Verfeinerung. Die angegebenen Webseiten stellen dabei eine Momentaufnahme des derzeitigen Entwicklungsstandes dar und dienen derzeit als State-of-the-art, was das Entwicklungsstadium betrifft.

# Anmerkung

## Interessenkonflikte

Keine angegeben.

# Literatur


1. Allison P. Fixed Effects Regression Methods for Longitudinal Data Using SAS. SAS Publishing; 2005.
2. Allison P. Missing Data. Thousand Oaks, CA: Sage; 2001.
3. Allison P. Multiple Imputation for Missing Data: A Cautionary Tale. Sociol Methods Res. 2000;28:301-9. DOI: 10.1177/0049124100028003003
4. Brodrecht K. Umsetzung verschiedener Ersetzungsmethoden von fehlenden Werten in SAS [Diplomarbeit]. Ulm: Hochschule Ulm, Medizinische Dokumentation und Informatik; 2005.

5. Deal K. Missing Something? Multiple imputation software might help find missing value data. Hamilton, Ontario: McMaster University; 2004. Available from: http://www.statsol.ie/documents/Ken_Deal_Missing_Something.pdf
6. Graham JW, Hofer SM. EMCOV reference manual. Los Angeles: Institute for Prevention Research, University of Southern California; 1993. Available from: http://ftp.cac.psu.edu/pub/people/jwg4/dos/emcov.txt
7. Harrell F, Alzola C. An Introduction to S and the HMISC and Design Libraries. Nashville: Vanderbilt University, School of Medicine; 2006. Available from: http://cran.r-project.org/doc/contrib/Alzola+Harrell-Hmisc-Design-Intro.pdf
8. Hohl K, Muche R, Brodrecht K, Ziegler C. Ersetzung fehlender Werte in SAS: zwei weiterentwickelte SAS-Makros. In: 10. Konferenz der SAS-Anwender in Forschung und Entwicklung. Aachen: Shaker Verlag; 2006.
9. Hohl K, Muche R, Ring C, Ziegler C. Fehlende Werte in der (Regressions-) Analyse von Datensätzen: zwei SAS-Makros. In: 9. Konferenz der SAS-Anwender in Forschung und Entwicklung. Aachen: Shaker Verlag; 2005. p. 99-108.
10. Hohl K. Umgang mit fehlenden Werten – Ersetzungsmethoden für fehlende Werte kategorialer Variablen in klinischen Datensätzen. Saarbrücken: Vdm Verlag Dr. Müller; 2008. p. 105-16.
11. Horton NJ, Lipsitz SR. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Values. Am Stat. 2001;55(3):244-54. DOI: 10.1198/000313001317098266
12. Hox JJ. A Review of Current Software for Handling Missing Data. Kwantitatieve Methoden. 1999;62:123-38.
13. King G, Honaker J, Joseph A, Scheve K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. Am Polit Sci Rev. 2001;95(1):49-69.
14. Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 1987.
15. Little RJA, Yau L. Intention-to-treat-Analysis for Longitudinal Studies with Drop-outs. Biometrics. 1996;52(4):1324-33. DOI: 10.2307/2532847
16. Molenberghs G, Kenward MG. Missing Data in Clinical Studies. Chichester: John Wiley & Sons; 2007.
17. Muche R, Ring C, Ziegler C. Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Aachen: Shaker Verlag; 2005.
18. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Surv Methodol. 2001;27(1):85-95.
19. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.
20. Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.
21. Schafer JL. Imputation of missing covariates under a multivariate linear mixed model [Technical Report]. Pennsylvania: Dep. of Statistics, Penn. State University; 1997. Available from: http://www.stat.psu.edu/reports/1997/tr9704.pdf
22. van Buuren S, Oudshoorn CGM. Multivariate Imputation by Chained Equations: MICE V1.0 User's manual. TNO Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid; 2000. Available from: http://www.stefvanbuuren.nl/publications/MICE%20V1.0%20Manual%20TNO000038%202000.pdf
23. van Buuren S. Mistress 1.17 documentation. Statistiekreeks 92/07. Leiden: NIPG-TNO; 1992.
24. Völkner T. Der Einfluss des Umgangs mit fehlenden Werten auf die Evaluation von Behandlungsffekten in Messwiederholungsdesigns [Diplomarbeit]. Freiburg: Universität Freiburg; 2005.
25. von Hippel PT. Biases in SPSS 12.0 Missing Value Analysis. Am Stat. 2004;58(2):160-4. DOI: 10.1198/0003130043204
26. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials. 2004;1(4):368-76. DOI: 10.1191/1740774504cn032oa
27. Yuan YC. Multiple Imputation for Missing Data: concepts and new development. Rockville MD: SAS Institute Inc.; 2000. Available from: http://support.sas.com/rnd/app/papers/multipleimputation.pdf [aufgerufen am 26.02.2009]



**Korrespondenzadresse:**
Rainer Muche
Institut für Biometrie, Universität Ulm, Schwabstrasse 13, 89075 Ulm, Deutschland
rainer.muche@uni-ulm.de



***Bitte zitieren als***
*Mayer B, Muche R, Hohl K. Software zur Behandlung und Ersetzung fehlender Werte. GMS Med Inform Biom Epidemiol. 2009;5(2):Doc15.*

***Artikel online frei zugänglich unter***
*http://www.egms.de/en/journals/mibe/2009-5/mibe000094.shtml*

***Veröffentlicht:*** *27.10.2009*