

Der Umgang mit fehlenden Werten in Verlaufsstudien im Falle von Drop-outs

Benjamin Mayer

Institut für Epidemiologie und Medizinische Biometrie

Schwabstraße 13

89075 Ulm

benjamin.mayer@uni-ulm.de

Zusammenfassung

Fehlende Werte sind nahezu allgegenwärtig in der Forschung. Eines der Hauptprobleme fehlender Daten stellt die mögliche Verzerrung der interessierenden Parameter dar, die aus den Daten geschätzt werden sollen. Dies kann dann zu einer Verfälschung und Einschränkung der Aussagekraft von Untersuchungsergebnissen führen. Es gibt verschiedene Ansätze dem Problem fehlender Werte zu begegnen. Dies wurde in dem Übersichtsvortrag von Frau Steindorf auf der 15. KSFE in Heidelberg sehr schön dargelegt. Insbesondere im Rahmen von Longitudinalstudien ist die vollständige Erfassung aller relevanten Daten komplex und kann daher kaum realisiert werden. Es bedarf dabei zusätzlich den Umgang mit so genannten Drop-outs, die einen Spezialfall fehlender Werte darstellen. In diesem Vortrag sollen verschiedene Ansätze von Ersetzungsmethoden speziell für die Situation von Drop-outs in Verlaufsstudien vorgestellt werden. Es wird diskutiert, welche Voraussetzungen für die korrekte Anwendung der einzelnen Methoden erfüllt sein müssen und wie sie im gegenseitigen Vergleich zu bewerten sind hinsichtlich Schätzvalidität und –genauigkeit.

Schlüsselworte: Fehlende Werte, Dropout, PROC MI

1 Einleitung

In nahezu allen klinischen und pharmazeutischen 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 (siehe Abbildung 1). Die CCA bringt jedoch eine ganze Reihe bedeutender Probleme mit sich, die ihre Verwendbarkeit mehr als in Frage stellen: Die Nichtberücksichtigung ganzer Beobachtungseinheiten kann dazu führen, dass die Fall-

zahl drastisch reduziert wird, die Variabilität der Merkmale sich verändert, die Aussagekraft der Studie vermindert wird und Parameterschätzer aufgrund der evtl. zerstörten Strukturgleichheit verzerrt sind. Darüber hinaus steht sie in Widerspruch zum Auswertungsprinzip der so genannten Intention-to-treat-Analyse (ITT), 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. [1,2]

Complete Case Analysis

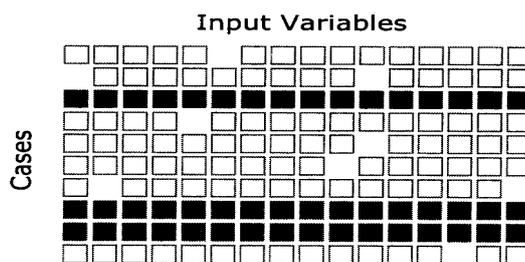


Abbildung 1: Complete Case Analyse

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 longitudinaler Daten ist die Durchführung einer CCA sehr problematisch, da hierfür jeder Patient, der mindestens einmal während der Verlaufsbeobachtung einen fehlenden Wert aufweist, für die Auswertung nicht berücksichtigt werden würde. Damit ist die Wahrscheinlichkeit, dass ein Patient von Fehlwerten betroffen ist, in Verlaufsstudien deutlich erhöht. Insbesondere sind dabei Dropouts, d.h. Studienabbrucher, die ab einem bestimmten Zeitpunkt keine Daten mehr aufweisen, ein Problem.

Dieser Beitrag soll eine Übersicht der zur Verfügung stehenden Möglichkeiten für die Behandlung von fehlenden Werten sein, die im Rahmen von Verlaufsstudien als Dropouts auftreten. Es soll zunächst ein Überblick zur Diagnostik fehlender Werte gegeben werden, bevor die wichtigsten Ersetzungsstrategien der Single Imputation (SI) und der Multiple Imputation (MI) vorgestellt werden. Den Abschluss bildet eine kurze Zusammenfassung.

menfassung der vorgestellten Methoden hinsichtlich ihrer Validität und Genauigkeit bei der Schätzung der Ersatzwerte.

2 Diagnostik und Ersetzungsstrategien

Die Aussagekraft von Studienergebnissen basierend auf einem Datensatz mit (ursprünglich) fehlenden Werten hängt stark von den Ergebnissen der so genannten **Missing Data Diagnostik (MDD)** ab. Ein Teil davon besteht aus der deskriptiven 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 Diagnostik beschreibt die Anordnung der fehlenden Werte im Datensatz, dem so genannten **Missing Data Pattern (MDP)**, und den (möglichen) Gründen für das Auftreten der fehlenden Werte, dem so genannten **Missing Data Mechanism (MDM)**. Letzteres ist wichtig für die Wahl einer geeigneten Ersetzungsmethode.

Bei der Bestimmung des Patterns unterscheidet man im Wesentlichen zwischen zwei Mustern. Fehlen die Werte breit gestreut und mehr oder weniger vereinzelt über den ganzen Datensatz hinweg (so genannte Studienaussetzer im Verlauf), so spricht man von einem beliebigen oder auch nicht-monotonen Muster. Im Gegensatz dazu steht ein monotonen 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 (siehe Abbildung 2). Diese Form des MDP beschreibt gerade die Situation, die sich aufgrund von Dropouts (so genannten Studienabbrechern) ergibt.

	Y ₁	Y ₂	Y ₃	Y ₄
x	x	x	.	.
x	x	x	.	.
x	x	.	.	.
x

(a) monotonen Muster

	Y ₁	Y ₂	Y ₃	Y ₄
x	.	x	x	.
x	x	x	.	.
x	x	.	x	.
x	x	x	.	.

(b) beliebiges Muster

Abbildung 2: Formen des Missing Data Pattern

Die drei verschiedenen Ausprägungen des Missing Data Mechanismus seien hier nur kurz erwähnt, für eine genauere Beschreibung siehe [3] oder auch [4]. Man unterscheidet drei Kategorien: Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR). Bei MCAR ist die Drop-out-Wahrscheinlichkeit, also die Wahrscheinlichkeit die Studie abzubrechen, in keinsten Weise

abhängig von bereits realisierten Werten der Zielgröße oder anderen Kovariablen des Datensatzes. MAR heißt, dass die Drop-out-Wahrscheinlichkeit nur von den bereits beobachteten Werten (der Zielvariablen oder Kovariablen) abhängt, wobei MNAR bedeutet, dass die Wahrscheinlichkeit für einen Drop-out zusätzlich von den fehlenden Werten selbst abhängt. Allerdings ist es nahezu unmöglich, den vorliegenden Mechanismus explizit zu identifizieren und in den realen Daten nachzuweisen.

Missing Data Mechanismus	Beschreibung
Missing Completely At Random	Das Auftreten eines fehlenden Wertes in der Variable Y ist nicht abhängig a) von der Ausprägung der Variable Y selbst oder b) den restlichen Variablen X_1 bis X_n im Datensatz
Missing At Random	Das Auftreten eines fehlenden Wertes in einer Variable Y ist vollständig durch die Ausprägungen der restlichen Variablen X_1 bis X_n erklärbar
Missing Not At Random	Das Auftreten von fehlenden Werten in der Variable Y ist a) von der (unbekannten) Ausprägung der Variable Y abhängig b) nicht durch die Ausprägungen der übrigen Variablen X_1 bis X_n erklärbar

Abbildung 3: Ausprägungen des Missing Data Mechanismus

Oftmals kann keine strikte Abgrenzung eines bestimmten Mechanismus vorgenommen werden, da es sich um eine Mischform handelt.

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 Imputation-Verfahren 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 [1,3].

Bei der **Multiple Imputation (MI)** wird ein fehlender Wert durch mehrere ($m > 1$) plausible Werte ersetzt, so dass m vervollständigte Datensätze aus der Ersetzung resultieren. Diese Datensätze werden einzeln mit der gleichen Methode, basierend auf einem jeweils kompletten Datensatz, ausgewertet. Anschließend werden die Ergebnisse dieser Analysen zu gemeinsamen Schätzern und Standardfehlern zusammengefasst. Das Vorgehen der MI ist in der folgenden Abbildung graphisch dargestellt und in Little & Rubin [1] genauer erläutert.

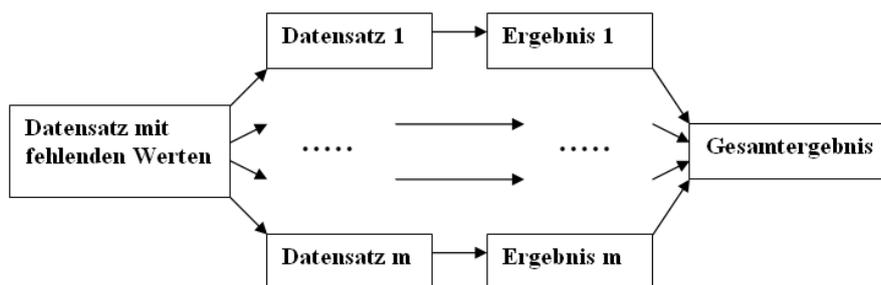


Abbildung 4: Schema der Multiple Imputation

Der entscheidende Vorteil der MI gegenüber den SI-Verfahren ist die korrekte Berücksichtigung des Standardfehlers. 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 mit beachtet. Demzufolge werden auch Konfidenzintervalle und p-Werte korrekter berechnet, als das mit einer beliebigen SI-Methode möglich wäre.

3 Der Umgang mit Dropouts in SAS

In diesem Kapitel werden die am häufigsten angewandten und in der Literatur erwähnten Ersetzungsmethoden für Drop-outs kurz vorgestellt und deren Umsetzbarkeit in SAS diskutiert. Für eine ausführliche Beschreibung und Diskussion der einzelnen Methoden siehe [5].

3.1 SI-Methoden

LOCF

Eine der populärsten SI-Methoden in der Praxis ist (noch immer) die Last-Observation Carried-Forward-Methode (LOCF), die trotz einiger Kritikpunkte häufig eingesetzt wird bei Longitudinalstudien, in denen Drop-outs vorliegen. Bei LOCF werden alle Missing Values eines Patienten durch den letzten beobachteten Wert, der für eben diesen Studienteilnehmer erfasst werden konnte, als „Schätzer“ für alle unbekanntenen Werte bis zum Beobachtungsende verwendet. Man kann nicht von einem guten Schätzer im statistischen Sinne sprechen, da der via LOCF generierte Ersetzungswert weder konsistent noch unverzerrt ist, was für die Güte eines Schätzers wichtige Kriterien sind. Der resultierende vervollständigte Datensatz ermöglicht es dem Untersucher dann zwar eine Datenanalyse wie gewohnt vorzunehmen, allerdings sollte bedacht werden, dass bei LOCF davon ausgegangen wird, dass alle unbeobachteten Zielgrößen dem letzten beobachteten Wert entsprechen. Wird diese Annahme verletzt, was im Allgemeinen sehr wahrscheinlich ist, hat dies eine verzerrte Schätzung des Mittelwerts bzw. der Varianz der Daten zur Folge, so dass die LOCF-Methode in ihrer Anwendung sehr beschränkt ist.

Das LOCF-Verfahren wird insbesondere für die schlechte Performance bei MCAR kritisiert. Zudem ist nicht klar, warum gerade der letzte beobachtete Wert eines Patienten ein plausibler Wert für alle fehlenden Werte dieses Patienten sein sollten. Ein Grund, weshalb die LOCF-Methode trotz den angesprochenen Punkten dennoch oftmals in der Praxis Anwendung findet, ist ihre in manchen Situationen konservative Eigenschaft. Neben den in manchen Situationen positiven konservativen Eigenschaften ist der Gebrauch von LOCF sicherlich auch ganz naiv darin begründet, dass man mit Hilfe dieser Methode auf einfache Weise einen vollständigen Datensatz zur Weiterverarbeitung erhält, ohne einen großen Aufwand zur Ersetzung der fehlenden Werte betreiben zu müssen.

Durch ein entsprechend konzipiertes SAS-Makro kann auf einfache Weise die LOCF-Methode umgesetzt werden. Die Bestimmung des letzten beobachteten Wertes je Pati-

ent kann über konventionelle DATA-Steps realisiert werden. Eine eigens für dieses Verfahren entwickelte SAS-Prozedur gibt es jedoch nicht.

Conditional Mean (Regression)

Die Idee dieses Ansatzes ist es, auf der Basis aller vollständig beobachteten Fälle (Zielgröße und Kovariablen) ein Regressionsmodell der Form

$$\mathbf{y}_{obs} = \mathbf{x}_{obs}\beta$$

aufzustellen. Hierbei markiert \mathbf{y}_{obs} den Vektor aller beobachteten Zielgrößen und \mathbf{x}_{obs} die entsprechend beobachteten Kovariablen. Die dabei berechneten Regressionskoeffizienten β können anschließend dafür verwendet werden, die fehlenden Zielgrößen zu schätzen. Voraussetzung hierfür ist jedoch, dass die Kovariablen für alle Patienten erfasst werden konnten.

Zur Umsetzung der Regressionsersatzung sind in SAS verschiedene Prozeduren verfügbar. In [5] wurde die MIXED-Prozedur verwendet, die sich mit der REPEATED Anweisung sehr gut für den Einsatz in longitudinalem Studiensetting eignet.

LOCF-Varianten

In der Literatur finden sich neben dem klassischen LOCF-Verfahren weitere Varianten desselben Prinzips. Diese Varianten arbeiten im Wesentlichen wie LOCF, jedoch dient nicht der letzte Beobachtungswert eines Patienten als Ersatzwert, sondern es werden andere Referenzwerte zur Ersetzung herangezogen. Bei der First-Observation-Carried-Forward (FOCF) wird der jeweils erste Beobachtungswert je Patient als Ersatzwert für die fehlenden Werte verwendet. Die Idee hinter diesem Ansatz scheint noch weniger intuitiv als bei LOCF: Während bei der Verwendung des letzten beobachteten Wertes eventuell Informationen über die Entwicklung der Zielgröße bis zu diesem Zeitpunkt berücksichtigt werden, kann der Baseline-Wert der Zielgröße sicherlich nur den Zustand vor Beginn der Studie wiedergeben.

Das Verfahren der Worst-Observation-Carried-Forward (WOCF) ist in der Literatur auch bekannt unter dem Namen "Worst-Case-Analyse". Dabei wird der schlechteste Messwert, den man jeweils für einen Patienten bisher beobachten konnte, als Ersatzwert verwendet. Man erreicht auf diese Weise eine konservative Schätzung der Zielgröße am Studienende. Die Best-Observation-Carried-Forward ist demnach ein Ansatz, bei dem der jeweils beste erfasste Zielgrößenwert wird für die Ersetzung der fehlenden Werte genutzt wird. Einer konservativen Schätzung der Effektgröße via WOCF-Methode steht in diesem Fall eine optimistische Schätzung gegenüber.

Die drei vorgestellten LOCF-Varianten werden häufig simultan angewandt, um auf Basis der erwartungsgemäß unterschiedlichen Ergebnisse eine Sensitivitätsanalyse durchführen zu können.

Wie für die LOCF-Methode existieren auch für die vorgestellten Varianten keine eigens in SAS entwickelten Prozeduren. Vielmehr muss auch für die Varianten auf die Verwendung geeigneter SAS-Makros verwiesen werden, die sich im Falle der FOCF-Methode analog zum LOCF-Makro leicht realisieren lassen. Individueller muss die Pro-

grammierung entsprechender Makros für die WOCF und BOCF Varianten realisiert werden, da die Entscheidung, was als „schlechtester“ bzw. „bester“ Wert jeweils anzusehen ist, unmittelbar von den Daten selbst abhängt.

EM-Algorithmus

Ein Verfahren zur Bestimmung von Maximum-Likelihood-Schätzern unter unvollständiger Information ist der so genannte Expectation-Maximization-Algorithmus (EM). Er geht auf Dempster, Laird und Rubin [6] zurück und beschreibt ein iteratives Verfahren zur Parameterschätzung in Fällen, in denen die betreffende Zielvariable fehlende Werte aufweist. Die Daten unterliegen dabei einer unbekanntem Verteilung, die durch einen ebenfalls unbekanntem Parameter θ bestimmt wird.

Im Expectation(E)-Step werden die fehlenden Elemente der Zielgröße auf Basis der beobachteten Zielgrößenwerte und des Parameterschätzers θ aus dem vorherigen Iterationsschritt geschätzt. Im Maximization(M)-Step wird dann mit Hilfe der Maximum-Likelihood-Methode ein (neuer) Schätzer für den Verteilungsparameter θ auf Basis der beobachteten und geschätzten Zielgrößenwerte berechnet. Der EM-Algorithmus bricht ab, wenn sich der Parameterschätzer θ im Vergleich zum voran gegangenen Iterationsschritt nur noch marginal verändert, d.h. ein bestimmtes Konvergenzkriterium erreicht ist. Ist der Parameterschätzer divergent, so endet der Iterationsprozess nach einer vorab festgelegten Zahl an Iterationsschritten. Zu Beginn des Verfahrens muss theoretisch ein Startwert für θ bestimmt werden, in der Praxis wird dieser jedoch zumeist standardmäßig von der verwendeten Software automatisch generiert. Der EM-Algorithmus seinerseits wird insbesondere zur Generierung von Startwerten für andere Ersetzungsverfahren, wie z.B. die im nächsten Abschnitt vorgestellte MCMC-Methode, verwendet und weniger zur tatsächlichen Ersetzung fehlender Werte.

Die SAS-Prozedur MI bietet den EM-Algorithmus als eine optionale Ersetzungsmethode an, die jedoch nicht standardmäßig von PROC MI ausgeführt wird.

3.2 MI-Methoden

MCMC

Die Markov Chain Monte Carlo-Methode (MCMC) nutzt die vorhandene Information der beobachteten Daten, um auf Basis verschiedener Algorithmen (z.B. Data Augmentation (DA) [7]) eine entsprechende a-posteriori-Verteilung zu berechnen, aus der dann Ziehungen für die Ersetzung der fehlenden Werte vorgenommen werden. Beim DA-Algorithmus handelt es sich um ein iteratives Verfahren, wobei oftmals die Ergebnisse des EM-Algorithmus als Startwerte benutzt werden. Die MCMC-Methode ist als sehr robustes Verfahren zur Imputation fehlender Werte von stetigem Messniveau bekannt und kann sowohl im Falle monotoner, als auch beliebiger Fehlwertmuster angewandt werden. Das Verfahren liefert selbst bei relativ hohen Fehlwertanteilen valide Ergebnisse unter MCAR bzw. MAR.

Die MCMC-Methode ist als Standardverfahren in PROC MI implementiert. Aufgrund ihres breiten Einsatzspektrums ist die Anwendung der Methode unabhängig vom vorliegenden MDP. Einzige Voraussetzung ist jedoch, dass es sich bei der Zielgröße, die von fehlenden Werten betroffen ist, um eine stetige Variable handelt. Mittels der MCMC-Methode können unvollständige Datensätze entweder komplett vervollständigt werden, oder nur soweit, dass ein monotonen MDP entsteht.

Stochastische Regression

Die im letzten Abschnitt vorgestellte Regressionsmethode wird additiv erweitert um einen zufälligen Fehlerterm. Es resultiert die Gleichung

$$\mathbf{y}_{obs} = \mathbf{x}_{obs}\beta + \epsilon$$

wobei ϵ der Vektor der Zufallsfehler ist. Auf diese Weise wird dem Ersatzwert bei jeder Durchführung des Verfahrens ein zufälliger Wert addiert, weshalb man bei Modell von so genannter stochastischer Regressionsersetzung spricht, die nun eine Multiple Imputation ermöglicht. Die Methode der stochastischen Regression liefert allerdings wiederum nur dann valide Ergebnisse, wenn die fehlenden Werte gemäß einem MAR- oder MCAR-Mechanismus im Datensatz vorliegen.

Das Verfahren der stochastischen Regressionsersetzung setzt zudem ein monotonen MDP voraus, so dass in PROC MI die REG Methode nur über die MONOTONE Anweisung erreicht werden kann.

Predictive Mean Matching

Die Methode des Predictive Mean Matching wurde ebenfalls für die Ersetzung fehlender Werte stetiger Zielgrößen konzipiert, wurde von Little [8] entwickelt und geht von denselben Voraussetzungen eines monotonen MDP, sowie eines MAR- oder MCAR-Mechanismus aus. Zusätzlich gelte für die Daten die Annahme einer multivariaten Normalverteilung. Sie ist im Wesentlichen sehr ähnlich zu dem besprochenen Regressionsverfahren, jedoch mit gewissen Unterschieden.

Der so genannte Predictive Mean (PM) einer Beobachtung bzw. Zielgröße ist gegeben durch

$$\hat{\mathbf{y}} = \mathbf{x} \hat{\beta}.$$

Für jeden fehlenden Zielgrößenwert wird gemäß der PM-Matching Methode dann zufällig ein Wert aus einer Menge M_0 von beobachteten Zielgrößenwerten gezogen und als Ersatzwert verwendet, deren PM möglichst ähnlich zu dem PM des fehlenden Wertes sein sollten. Die prognostizierten Werte entstammen dabei dem o.g. Regressionsmodell, weshalb auch für die PM Matching Methode die Annahme vollständiger Kovariablen getroffen wird. Für eine kleine Menge M_0 von beobachteten Referenzwerten sind höhere Korrelationen zwischen den multiplen Imputationen der Fehlwerte zu erwarten und somit eine höhere Variabilität der Punktschätzer. Andererseits bewirkt ein größeres M_0 einen weniger starken Einfluss des Imputationsmodells und verzerrte Parameterschätzer. Die Predictive Mean Matching Methode liefert selbst bei geringen Abweichungen der Normalverteilungsannahme plausible Ersatzwerte.

Aufgrund der Voraussetzung für das Verfahren, dass die fehlenden Werte gemäß einem monotonen Muster angeordnet sein müssen, kann das Predictive Mean Matching in PROC MI nur unter der MONOTONE Anweisung mit REGPMM aufgerufen werden.

Propensity Score

Eine weitere Ersetzungsmethode im Rahmen der MI stellt die Propensity Score Methode (PS) dar, welche für die Ersetzung stetiger und qualitativer Variablen konzipiert wurde und ein monotonen Fehlwertmuster im Datensatz voraussetzt. Im Kontext von fehlenden Werten kann ein PS interpretiert werden als die bedingte Wahrscheinlichkeit, dass ein Patient unter gegebener Kovariablenstruktur einen Fehlwert hat [9]. Auf Basis eines logistischen Regressionsmodells wird für jeden Fehlwert der betreffenden Variablen ein PS berechnet zur Schätzung der Wahrscheinlichkeit, dass der betreffende Wert fehlt. Alle Patienten des Datensatzes werden dann entsprechend der berechneten PS in aufsteigender Reihenfolge sortiert und in eine bestimmte Anzahl von Teilmengen eingeordnet. Die Ersetzung der Fehlwerte erfolgt dann über die zufällige Ziehung eines beobachteten Wertes aus der entsprechenden Teilmenge. Bedeutende Restriktionen für das Verfahren stellen unvollständige Kovariablen und ein NMAR Mechanismus dar. Wiederum gilt im Rahmen der Umsetzung mit der SAS-Prozedur MI, dass der entsprechende PROPENSITY-Aufruf in der MONOTONE Anweisung erfolgt.

Discriminant Function

Die Discriminant Function Method (DFM) ist speziell für die Ersetzung von nominalen Variablen geeignet. Sie setzt MAR und ein monotonen MDP voraus. Ferner wird angenommen, dass die Kovariablen im Modell der Diskriminanzanalyse normalverteilt sind [3].

Die Information über eine nominale Zielgröße Y mit fehlenden Werten und Ausprägungen $1, \dots, k$, sowie normalverteilten Kovariablen und ungefähr gleichen Kovarianzen wird für die Erstellung linearer Diskriminanzfunktionen verwendet. Auf Basis dieser Funktionen und entsprechenden a priori-Wahrscheinlichkeiten für die Zugehörigkeit eines Zielgrößenwertes zu einer bestimmten Klasse werden die a posteriori-Wahrscheinlichkeiten für das Auftreten der jeweiligen Ausprägung von Y berechnet. Für diese Wahrscheinlichkeiten gilt $p_1 + \dots + p_k = 1$. Zudem wird eine zwischen 0 und 1 gleichverteilte Zufallsvariable u generiert. Für $u < p_1$ wird der fehlende Wert der Beobachtung durch die erste Ausprägung ersetzt. Für $p_1 < u < p_1 + p_2$ entspricht der ersetzte Wert der 2. Ausprägung von Y usw. Ein entscheidender Nachteil bei der DFM ist, dass nur stetige normalverteilte Kovariablen berücksichtigt werden können.

In PROC MI kann die Methode unter der MONOTONE Anweisung mit DISCRIM aufgerufen werden.

Logistic Regression

Das Verfahren der logistischen Regression eignet sich unter MCAR bzw. MAR zur Ersetzung fehlender Werte für binäre und ordinalskalierte Variablen, wobei ebenfalls ein

monotones Fehlwertmuster (Drop-outs) vorausgesetzt wird. Es wird ein Regressionsmodell auf Basis der vollständig beobachteten Fälle aufgestellt, aufgrund dessen eine Schätzung der Regressionsparameter möglich ist, so dass anschließend auf dieser Basis eine Schätzung der fehlenden Werte vorgenommen werden kann. Dabei wird für die gegebenen Kovariablen die Wahrscheinlichkeit p berechnet ist, dass einer der beiden möglichen Werte 1 oder 2 an Stelle der fehlenden Werte der Variablen angenommen wird. Zusätzlich wird eine gleichverteilte Zufallsvariable u zwischen 0 und 1 berechnet, mit dieser dann die zuvor berechnete Wahrscheinlichkeit verglichen wird. Für u kleiner p wird dann eine 1 als Ersatzwert generiert, andernfalls eine 2.

Da auch diese Methode auf einem monotonen MDP basiert, kann der LOGISTIC-Aufruf in PROC MI nur unter der MONOTONE Anweisung erfolgen.

3.3 Methodenvergleich

In [5] wurden verschiedene Ersetzungsmethoden in longitudinale Studiendesign im Rahmen einer Simulationsstudie miteinander verglichen. Dabei wurde eine stetige Zielgröße untersucht in verschiedenen Missing Data Szenarien (unterschiedliche Mechanismen (MCAR, MAR, MNAR) und unterschiedliche Fehlwertanteile im Datensatz (5% - 50%)). Die im Folgenden präsentierten Ergebnisse decken also nicht diejenigen Methoden ab, die für qualitative Zielgrößen konzipiert wurden. Für die Simulation wurden fünf verschiedene Verfahren verglichen hinsichtlich Schätzgenauigkeit und -validität: LOCF, FOCF, Conditional Mean (Regression), MCMC und stochastische Regression.

Zur Bewertung der Ersetzungsgüte wurde jeweils die Differenz zwischen originalen und ersetzten Werten, gemittelt über alle Patienten des Datensatzes und die Anzahl aller Simulationsdurchläufe hinweg, berechnet.

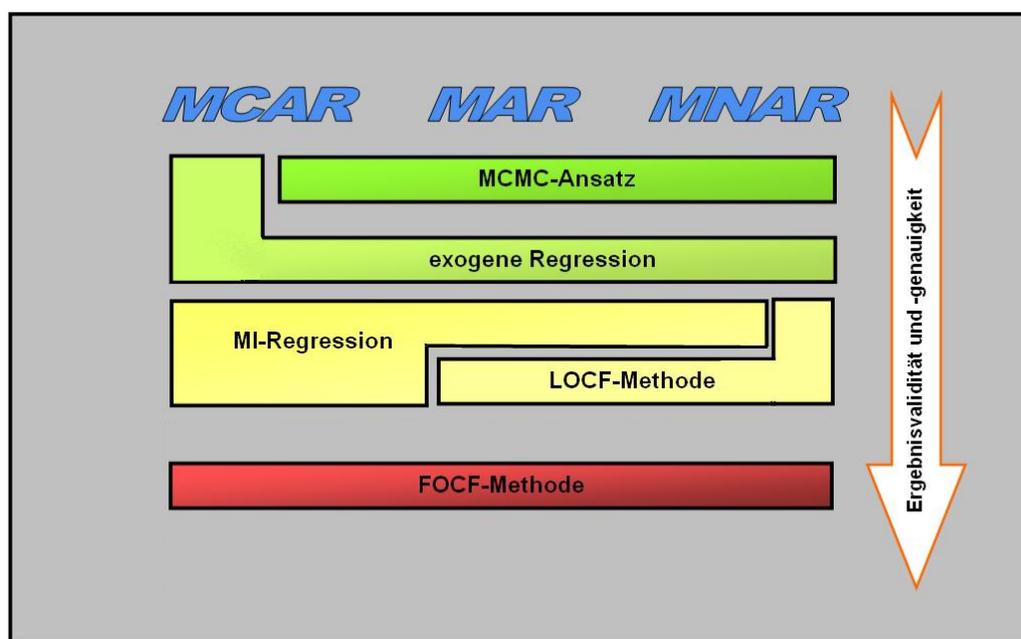


Abbildung 5: Methodenvergleich im Longitudinaldesign

Insgesamt zeigte die MCMC-Methode der MI-Strategie die besten Ergebnisse im Sinne von Genauigkeit und Validität der Ersatzwertschätzung. Die LOCF- und FOCF-Verfahren schlossen den Vergleich am schlechtesten ab, während die beiden Regressionsansätze (SI und MI) mit deutlichem Abstand zu MCMC folgen.

4 Zusammenfassung

Die bekannte Problematik fehlender Werte wird im Rahmen von klinischen Verlaufsstudien durch potentielle Studienabbrecher zusätzlich verstärkt. Drop-outs können unter Umständen zu einer massiven Einschränkung der auswertbaren Fälle führen. Umso wichtiger ist es daher, die Fehlwerte durch eine angemessene Methodik zu ersetzen. Speziell für die Situation eines monotonen Fehlwertmusters, welches durch Drop-outs erzeugt wird, existieren zahlreiche Methoden, die sich auch problemlos in SAS umsetzen lassen.

Dieser Beitrag stellte verschiedene Ersetzungsmethoden für fehlende Werte in Verlaufsstudien vor. Deren Umsetzung in SAS basiert im Wesentlichen auf der SAS-Prozedur PROC MI, welche fast alle gängigen Ersetzungsmethoden beinhaltet. PROC MI bietet dabei die Möglichkeit, fehlende Werte verschiedener Skalenniveaus zu ersetzen durch die Anwendung verschiedener verfügbarer Optionen. Für die Anwendung weiterer ad hoc Ersetzungsmethoden, wie beispielsweise der Last-Observation-Carried-Forward (LOCF), gibt es in SAS keine implementierten Prozeduren. Allerdings eignet sich die Syntax-Sprache in SAS sehr gut, um die entsprechenden Ansätze in Form von SAS-Makros umzusetzen.

Insgesamt deckt PROC MI als Standardprozedur in SAS zur Ersetzung fehlender Werte mittlerweile ein sehr breites Spektrum ab, so dass eine universelle Grundbehandlung fehlender Werte mit PROC MI durchaus empfohlen werden kann. Die Prozedur wird dabei standardmäßig mit dem MCMC-Ansatz (Markov Chain Monte Carlo) ausgeführt, welcher eine Ersetzung auch im Falle von Drop-outs ermöglicht.

Literatur

- [1] Little R.J.A., Rubin D.B. (1987) *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York
- [2] Wood A.M., White I.R., Thompson, S.G. (2004) Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*. 1: 368-376
- [3] Hohl K. (2008) Umgang mit fehlenden Werten – Ersetzungsmethoden für fehlende Werte kategorialer Variablen in klinischen Datensätzen. Vdm Verlag Dr. Müller, Saarbrücken, Seite 105-116
- [4] Molenberghs G., Kenward M.G. (2007) *Missing Data in Clinical Studies*. J. Wiley & Sons, Chichester

- [5] Mayer B. (2011) Fehlende Werte in klinischen Verlaufsstudien – Der Umgang mit Studienabbruchern. Dissertation an der Universität Ulm
- [6] Dempster A.P., Laird R., Rubin D.B. (1977) Maximum Likelihood for Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* 39: 1-38
- [7] Tanner M.A., Wong W.H. (1987) The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* 18: 528-540
- [8] Little R.J.A. (1988) Missing Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics* 6: 287-296
- [9] Rosenbaum P.R., Rubin D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55