Multiple Imputation – der State-of-the-Art-Umgang mit fehlenden Werten

Karen Steindorf
Deutsches Krebsforschungszentrum,
AG Umweltepidemiologie (C030)
und Abteilung für Präventive Onkologie
Im Neuenheimer Feld 280
69120 Heidelberg
k.steindorf@dkfz.de

Oliver Kuß
Institut für Medizinische Epidemiologie,
Biometrie und Informatik
Medizinische Fakultät,
Martin-Luther-Universität Halle-Wittenberg
Magdeburger Str. 8
06097 Halle (Saale)
oliver.kuss@medizin.uni-halle.de

Zusammenfassung

Fehlende Werte sind ein alltägliches Problem in der angewandten Forschung. In diesem Beitrag werden zunächst verschiedene vielfach eingesetzte Verfahren vorgestellt und ihre Praxistauglichkeit diskutiert. Es zeigt sich, dass die meisten dieser Verfahren unzureichende statistische Eigenschaften aufweisen. Kernelement des Beitrags ist daher das Verfahren der Multiplen Imputation, ein komplexes Verfahren zur Ersetzung von fehlenden Werten. Trotz seiner guten statistischen Eigenschaften und der mittlerweile verbesserten Verfügbarkeit von statistischer Standard-Software, in SAS zum Beispiel mittels PROC MI und PROC MIANALYZE, findet diese Methode bislang nur geringen Einsatz in der Praxis. Eine breitere Anwendung ist wünschenswert und wird sicherlich noch weitere praktische Fragen aufwerfen. Der Einsatz von Multipler Imputation erfordert statistische Fachkompetenz und ist sowohl bei der personellen aber auch bei der zeitlichen Projektplanung hinreichend zu berücksichtigen.

Schlüsselwörter: Fehlende Werte, Multiple Imputation, PROC MI, PROC MIANALYZE

1 Einleitung

Fehlende Werte sind ein alltägliches Problem in allen Bereichen, in denen große Datenmengen gesammelt und verarbeitet werden. Somit ist jede Auswertung, die den Umgang mit fehlenden Werten nicht angemessen adressiert, problematisch (Allison, 2001). Werte fehlen aus den verschiedensten Gründen, ein Interview wird abgebrochen, Teilnehmer verweigern die weitere Teilnahme in einer Langzeitbeobachtung, Fragen werden übersehen, Informationen sind nicht mehr verfügbar etc.

Zahlreiche Verfahren zur Ersetzung von fehlenden Werten wurden in der Vergangenheit vorgeschlagen und diskutiert (Little & Rubin, 1987; Schafer, 1997). In neuerer Zeit zeigte sich zunehmend, dass vor allem aufwändigere Verfahren adäquate Vorgehens-

weisen darstellen (Donders et al., 2006; Schafer & Graham, 2002). In der Praxis setzen sich diese jedoch nur langsam durch, obwohl mittlerweile viele statistische Auswertungspakete, u.a. SAS, eine computertechnische Umsetzung anbieten. In diesem Beitrag werden übersichtsartig einige der klassisch verwendeten Ansätze vorgestellt und ihre Praxistauglichkeit kurz erläutert, bevor das Verfahren der Multiplen Imputation präsentiert wird. Es wird sich zeigen, dass über die adäquate Methodik hinaus auch die gute Dokumentation zum Umgang mit fehlenden Werten in Berichten und Publikationen der angewandten Forschung zunehmend an Bedeutung gewinnt.

2 Ad hoc Methoden

Es gibt seit vielen Jahren verschiedene Ansätze, die für den Umgang mit fehlenden Werten vorgeschlagen und auch vielfältig eingesetzt werden.

2.1 Restriktions-Methoden

Ein Ansatz, der den Restriktions-Methoden zu zuordnen ist, ist die „Complete-Case Analyse“. Dabei werden nur Beobachtungen bei der Auswertung berücksichtigt, die in allen Variablen vollständig sind. In vielen SAS-Prozeduren, wie zum Beispiel PROC REG, entspricht dieses Vorgehen der Standardeinstellung. Vorteilhaft ist, dass diese Methode für jede Analyseart und ohne zusätzliche Software verwendet werden kann. Ein Problem liegt in der Nicht-Verwendung von erhobenen Daten, da die komplette Datenzeile mit allen anderen vorhandenen Informationen aufgrund des fehlenden Wertes einer oder weniger Variablen ausgeschlossen wird. Dies kann finanziell und auch ethisch unbefriedigend sein. Diese Methode verringert die Effizienz der statistischen Verfahren, d.h. es führt zu verringerter Power der statistischen Tests bzw. zu breiteren Konfidenzintervallen. Als intuitiv besseres Verfahren wurde daher vorgeschlagen, für verschiedene Auswertungen zu einem Datensatz, die Daten nur bezüglich der Variablen einzuschränken, die gerade relevant sind, d.h. wenn die Variablen X,Y und Z erhoben wurden und die Korrelation zwischen X und Y berechnet werden soll, dass man in diesem Fall Beobachtungen mit fehlenden Werten in der Variablen Z durchaus berücksichtigen könne. Neben der problematischen Berichterstattung der Ergebnisse mit variierenden Fallzahlen wurde gezeigt, dass dieses Verfahren zu stark verzerrten und ineffizienten Schätzern führen kann (Allison, 2001). Dieser Ansatz ist daher für die Praxis nicht zu empfehlen.

2.2 Einfache Substitutions-Methoden

Die grundlegende Idee dieser Methoden ist es, die fehlenden Werte durch einen „vernünftigen“ Ersatzwert zu ersetzen und dann die Daten mit dem vervollständigten Datensatz wie gewohnt auszuwerten. Die Ersatzwerte werden üblicherweise aus den vollständig vorliegenden Datensätzen gewonnen, indem allgemeine Mittelwerte, Mittelwerte in passenden Untergruppen oder die Ergebnisse aus Regressionsanalysen verwendet werden. Das allgemeine Problem dieser Verfahren besteht darin, dass die weiteren Analy-

sen vernachlässigen, dass fehlende Werte ersetzt wurden. Die mit der Ersetzung verbundene statistische Unsicherheit bleibt somit unberücksichtigt, Standardfehler werden systematisch unterschätzt und die statistische Inferenz ist fehlerbehaftet.

Eine alternative Methode ist es, allen fehlenden Werten eine eigene Kategorie zu zuordnen, so dass diese als eigener Datenwert in die Analyse einfließen kann oder über eine Indikatorvariable in den Modellen berücksichtigt werden kann. Diese Ansätze sind in der Epidemiologie derzeit weit verbreitet. Insbesondere die Verwendung einer gesonderten Kategorie hat gerade für die Deskription der Daten viele Vorteile.

2.3 Vergleich der verschiedenen Ad-hoc Methoden

Eine vertiefende Bewertung der statistischen Eigenschaften der verschiedenen Verfahren setzt voraus, dass auch die Entstehungsmechanismen der fehlenden Werte berücksichtigt werden. So unterscheidet man die Fälle, in denen die vollständigen Beobachtungen eine Zufallsstichprobe aller Beobachtungseinheiten darstellen (sogenanntes MCAR=„Missing completely at random“) oder in denen die Wahrscheinlichkeit für eine fehlende Beobachtung in einer Variablen Y unabhängig vom konkreten Wert von Y ist, nachdem für andere Variablen in der Analyse adjustiert wurde (MAR=„Missing at random, P(Wert von Y fehlt / Y, X) = P(Wert von Y fehlt / X)). Falls die Werte nicht MAR sind, so wird der Mechanismus als informativ (informative, not ignorable, MNAR=„Missing not at random“) bezeichnet.

In der Tat ist das Verständnis des zu Grunde liegenden Mechanismus, der die fehlenden Werte generiert hat, essentiell für deren Behandlung. So bestimmt der Mechanismus den Grad der Verzerrung und, zum Teil, die korrekte Wahl zum Umgang mit den fehlenden Werten. Eine Vertiefung dieser Aspekte ist an dieser Stelle nicht möglich. Es zeigte sich jedoch, dass von all diesen Verfahren die Complete-Case Analyse die beste ist. Gerade für die in der Epidemiologie häufig verwendete Logistische Regression konnte gezeigt werden, dass sie unter recht allgemeinen Bedingungen zu valider statistischer Inferenz führen kann (Vach, 1994). Dennoch bedeutet es, dass bei einer Untersuchung von 1000 Personen, bei denen 20 Variablen erhoben werden, die jeweils eine Fehlwahrscheinlichkeit von 5% aufweisen, am Ende nur 360 komplette Datensätze verwendet werden können, wohingegen die Angaben von 640 Personen nicht beachtet werden. Es erscheint somit sinnvoll, sich auch mit anderen Verfahren für fehlende Werte auseinander zu setzen, auch wenn sie unter Umständen deutlich aufwändiger sind.

Tabelle 1: Überblick über ausgewählte Verfahren zum Umgang mit fehlenden Werten

Ad hoc Methoden	**Verfeinerte Substitutions-Methoden**
(1) **Restriktionsverfahren** (2) **Einfache Substitutions-verfahren**	(1) **Maximum Likelihood Schätzung** (2) **Multiple Imputation** (3) **Pseudo Maximum Likelihood Schätzung**
• Schlechte statistische Eigenschaften (inkonsistente oder verzerrte Schätzer; Varianzschätzer und somit Tests und Konfidenzintervalle nicht valide) • Complete-case Analyse ist von diesen Verfahren in der Regel am besten. • Die anderen sind nicht zu empfehlen, einige sind ok unter der MCAR Annahme	• Methoden haben wünschenswerte statistische Eigenschaften, erfordern aber erhöhten Aufwand. • Die Verfügbarkeit von statistischer Standard-Software hat sich in den letzten Jahren dramatisch verbessert, z.B. in SAS. • Diese Methoden sollten verwendet werden, wenn eine größere Anzahl von Daten fehlt.

3 Das Verfahren der Multiplen Imputation

Es gibt verschiedene Verfahren, die bessere statistische Eigenschaften als die genannten Ad-hoc Methoden besitzen (s. Übersicht in Tabelle 1). Diese liefern sowohl unter der MCAR- als auch unter der MAR-Annahme approximativ unverzerrte und effiziente Schätzer. Damit bilden sie die Basis für valide statistische Inferenz, d.h. die Konfidenzintervalle halten die Überdeckungswahrscheinlichkeiten und die Tests die vorbestimmten Signifikanzniveaus ein. In diesem Beitrag wird nur auf das Verfahren der Multiplen Imputation eingegangen und dieses in seinen Grundzügen vorgestellt (Donders et al., 2006; Schafer & Graham, 2002; Spratt et al., 2010). Die anderen Ansätze werden nicht behandelt, liefern häufig aber sehr ähnliche Ergebnisse.

Die grundlegende Idee der Multiplen Imputation ist es, dass die Ersetzung eines fehlenden Wertes durch einen einzelnen Wert die mit der Ersetzung verbundene Unsicherheit nicht widerspiegeln kann. Daher wird ein einzelner fehlender Wert durch viele plausible Werte ersetzt, deren Variabilität die Unsicherheit darüber wiedergibt, welches der richtige Wert ist. Wie in Tabelle 2 dargestellt, gliedert sich das Verfahren der Multiplen Imputation insgesamt in drei Phasen.

Tabelle 2: Überblick über die drei wesentlichen Phasen einer Multiplen Imputation

Phasen	**Aufgabe**	**SAS-Prozedur**
Schritt 1	Die fehlenden Daten werden *m*-fach ersetzt, so dass *m* komplette Datensätze entstehen	PROC MI
Schritt 2	Die *m* kompletten Datensätze werden mit adäquaten statistischen Verfahren ausgewertet	Zum Beispiel: PROC LOGISTIC; PROC REG; PROC PHREG; PROC GLM
Schritt 3	Die Ergebnisse der *m* kompletten Datensätze werden zu einem Gesamtergebnis zusammengefasst	PROC MIANALYZE

In dem ersten Schritt wird, im Unterschied zu den Ad-Hoc-Substitutionsverfahren, nicht nur ein vollständiger, sondern *m* vollständige Datensätze erstellt. In der Literatur wurde bis vor kurzem empfohlen, 5 bis 10 komplette Datensätze zu erstellen, eine neuere Arbeit zeigt jedoch, dass auch größere Anzahlen erforderlich sein können (Spratt et al., 2010). Für den Fall, dass der einfache Datensatz bereits umfangreich ist, so kann diese Vervielfachung in der Praxis durchaus die Grenzen der Praktikabilität erreichen. Die Vervielfachung und die gleichzeitige Ersetzung der fehlenden Werte werden in SAS mit der Prozedur PROC MI umgesetzt. Für diesen Schritt ist die wichtige Unterscheidung zwischen dem Imputationsmodell und dem analytischen Modell zu treffen. Das Imputationsmodell beinhaltet alle Variablen des Datensatzes, in denen fehlende Werte aufgetreten sind, und alle Variablen, die einen Beitrag zum „Auffüllen" der fehlenden Werte liefern können. Das analytische Modell enthält alle Variablen, die für die geplante Auswertung erforderlich sind. Das Imputationsmodell ist somit mindestens so groß wie das analytische Modell, es kann jedoch auch weitere Variablen beinhalten, die die Qualität der Imputation verbessern können. Ob eine Variable Z ein Prädiktionspotenzial für die Ersetzung der fehlenden Werte der Variablen M haben kann, kann man unter anderem untersuchen, indem man den Anteil der fehlenden Werte in M für die verschiedenen Ausprägungen von Z betrachtet.

Die allgemeine Syntax in SAS sieht für diesen ersten Schritt der Multiplen Imputation mittels der SAS-Prozedur PROC MI wie folgt aus:

```
PROC MI < Optionen> ;
    BY Variablen ;
    CLASS Variablen ;
    EM < Optionen > ;
    FREQ Variablen ;
    MCMC < Optionen > ;
    MONOTONE < Optionen > ;
    TRANSFORM Transform ( Variablen < / Optionen >)
      < ... Transform ( Variablen < / Optionen >) > ;
    VAR Variablen ;
```

Ein grundlegendes Beispiel könnte wie folgt aussehen:

```
PROC MI data=KSFE_BSP seed=501213 out=OUT_KSFE;
     VAR  muscle_strength  oxygen  weight  height  gender  runtime
          fitness;
     RUN;
```

In diesem Beispiel wird mittels der Spezifikation des Startwertes des Zufallszahlengenerators (seed=) die Reproduzierbarkeit der Simulationsprozedur gewährleistet. Wichtig ist auch die Benennung eines Datensatzes, in den das Ergebnis ausgelesen wird. Nur so ist eine Weiterverarbeitung der Ergebnisse im nächsten Schritt möglich. In diesem sehr einfachen Beispiel wurden nur wenige der in SAS vorhandenen Spezifikationsmöglichkeiten genutzt, die anderen bleiben somit weiterhin mit Defaults belegt, von denen einige in Tabelle 3 zusammengestellt sind. Weitere Optionen dienen zum Beispiel der Festlegung von zugelassenen Wertebereichen (Minimum und Maximum) oder der Genauigkeit (round) der zu ersetzenden Werte.

Tabelle 3: Überblick über einige Default-Einstellungen der SAS-Prozedur PROC MI

Die MI Prozedur

Model Information	
Data Set	WORK.KSFE_BSP
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	501213

Als statistische Methodik zur Schätzung der fehlenden Werte wird als Standardeinstellung die Markov Chain Monte Carlo (MCMC) Methode angewendet, die von einer multivariaten Normalverteilung aller Variablen im Imputationsmodell ausgeht. Es gibt aber auch andere Ansätze, wie zum Beispiel nicht-parametrische oder Propensity Score-Verfahren. Wie bei allen auf Simulation beruhenden Methoden ist es in diesem Schritt überaus wichtig, die Konvergenz und Stabilität des Simulationsverlaufs zu überprüfen. Die verfügbaren Softwarepakete bieten hier Unterstützung an. Generell erfordert dieser Schritt gute statistische Expertise. Der Aufwand zum Beispiel schon für die Auswahl des Imputationsmodells aber auch für die Überprüfung der Stabilität der Ergebnisse ist nicht zu unterschätzen und in der Planung der Auswertungsphase sowohl personell als auch zeitlich zu berücksichtigen.

Im zweiten Schritt werden dann die *m* kompletten Datensätze, möglicherweise reduziert auf die Variablen, die nur zum Analysedatensatz gehören, mit den geplanten Analyse-

modellen und –verfahren einzeln ausgewertet. Dieser Schritt unterscheidet sich nicht von der Analyse, die mit dem Originaldatensatz durchgeführt worden wäre und kann jegliches statistisches Modell beinhalten. Insgesamt entsteht somit ein *m*-facher Ergebnisdatensatz. Eine Identifikationsvariable für die Zuordnung zu den *m* Datensätzen ist mit zu führen. Der Ergebnisdatensatz wird im dritten Schritt, z.B. mit der SAS Prozedur PROC MIANALYZE gemäß den Regeln von Rubin zu einem Gesamtergebnis zusammengefasst, das die Unsicherheiten über die *m* Wiederholungen der Auswertungen bei der statistischen Inferenz berücksichtigt. Auch hier unterstützen die Softwarelösungen die Klärung, wie viel Unsicherheit die Ersetzung der fehlenden Werte in die Ergebnisse gebracht hat.

4 Diskussion und Schlussfolgerungen

Der adäquate Umgang mit fehlenden Werten ist in zahlreichen Anwendungen nicht trivial, aber unabdingbar. Von den Ad-hoc-Verfahren erscheint derzeit die Methode der Complete-Case-Analyse am besten, trotz der erheblichen Datenverluste, die damit einhergehen können. Verfahren wie die Multiple Imputation weisen gravierende Vorteile gegenüber den derzeit als Standard verwendeten Methoden auf, so dass deren vermehrter Einsatz in der Praxis zu fordern ist. Die Zahl der verfügbaren statistischen Software-Lösungen hat in den letzten Jahren stetig zu genommen und bietet nun auf vielen Plattformen geeignete Lösungen an. Dennoch oder gerade für diese Ansätze geht der angemessene Umgang mit fehlenden Werten jedoch mit einem substantiellen Investment an Zeit und Energie einher. Bei der Planung und Durchführung von großen Datenerhebungen ist es daher von besonderer Bedeutung und unter Umständen auch ökonomisch sinnvoll, fehlende Werte so weit wie möglich zu vermeiden. Die zur Verfügung stehenden statistischen Methoden sollten insbesondere nicht dazu verleiten, diese Sorgfalt während der Studiendurchführung einzuschränken.

Unter der Erwartung, dass die verfeinerten und aufwändigeren Substitutions-Methoden in den nächsten Jahren vermehrt bei der Auswertung verwendet werden, ist auch eine genauere Untersuchung ihrer Praxistauglichkeit in verschiedenen Bereichen erstrebenswert. So wirft die Definition des Imputationsmodells in der Praxis häufig noch viele Fragen auf, deren Konsequenzen noch unzureichend verstanden sind. In jedem Fall ist zu beachten, dass alle Variablen, die bei der Modellierung eine Rolle spielen können, vorab auch im Imputationsmodell enthalten sein müssen. Das umfasst zum Beispiel auch alle Interaktionsterme und transformierte Variablen. In Anwendungen, in denen die Auswahl der relevanten Variablen erst im Verlauf der Analysen vorgenommen wird, wie zum Beispiel in einigen epidemiologischen Anwendungen, ist die Definition des Imputationsmodells somit keineswegs trivial und die Konsequenzen sicherlich auch noch nicht hinreichend statistisch untersucht.

Auch wenn die Fortschritte auf dem Gebiet der multiplen Imputation in den letzten Jahren immens waren, sei noch auf einige Schwierigkeiten hingewiesen. Immer noch unbefriedigend, sowohl von der theoretischen Seite als auch bei der Umsetzung in PROC

MI, ist zum einen der Umgang mit fehlenden Werten für kategoriale Variablen. Zum anderen liefert auch die multiple Imputation nur gültige Schätzer, wenn mindestens die MAR-Annahme erfüllt ist. Beim Vorliegen von „informative dropout" kann auch die Methode der multiplen Imputation nicht mehr angewandt werden. In diesem Falle muss der Dropout-Prozess mit Hilfe von Selektions- oder Mischungsmodellen explizit modelliert werden. Keines dieser Verfahren ist bisher automatisiert in SAS umgesetzt.

Zu beachten ist aber auch, dass nicht jeder fehlende Wert in einer Datenbank ein Problem darstellt, sondern dass es häufig auch „korrekte" fehlende Werte gibt. Ein Beispiel ist das Alter bei der ersten Geburt bei einem männlichen Studienteilnehmer oder die fehlende Anzahl von gerauchten Zigaretten bei einem/r Nichtraucher/-in oder die Ergebnisdatei zu einem Lateinischen Quadrat oder einem Two-Stage-Studiendesign, die *per se* schon nicht auf Vollständigkeit angelegt sind. Der inhaltliche Bezug zu den fehlenden Werten darf bei der Wahl der adäquaten Methoden nicht verloren gehen. „Korrekte" fehlende Werte sind explizit vor der multiplen Imputation als automatischer Prozedur zu schützen.

Ein wichtiger Punkt beim Umgang mit fehlenden Werten stellt auch die umfassende Berichterstattung dar. Nur so lassen sich die berichteten Ergebnisse hinreichend bewerten und qualitativ einordnen. Es stehen Richtlinien zur Verfügung, wie der Umgang mit fehlenden Werten in Berichten und Publikationen dokumentiert werden kann, die vermehrt zur Anwendung kommen sollten (Sterne et al., 2009).

Literatur

[1] Allison P (2001): Missing data. Sage Publications, Thousands Oaks.

[2] Donders AR, van der Heijden GJ, Stijnen T, Moons KG (2006): A gentle introduction to imputation of missing values. J Clin Epidemiology, 59, 1087-91.

[3] Little R, Rubin D (1987): Statistical analysis with missing data. Wiley, New York.

[4] Schafer JL (1997): Analysis of incomplete multivariate data. Chapman & Hall, London.

[5] Schafer JL, Graham JW (2002): Missing data: our view of the state of the art. Psychol Methods, 7(2), 147-77.

[6] Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, Tilling K (2010): Strategies for Multiple Imputation in Longitudinal Studies. Am J Epidemiol, 172, 478-87.

[7] Sterne JA, White IR, Carlin JB, Spratt M, Royston P et al. (2009): Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.

[8] Vach W (1994): Logistic regression with missing values in the covariates. Springer Verlag, Heidelberg.