Die Kombination der Ersetzung fehlender Werte und Variablenselektionsmethoden bei der Entwicklung von Prognosemodellen und deren Umsetzung mit SAS

Annabel Sandra Stierlin
Benjamin Mayer
Rainer Muche
Institut für Epidemiologie und Medizinische Biometrie
Uni Ulm
Schwabstraße 13
89075 Ulm
Annabel.Stierlin@Uni-Ulm.de

Zusammenfassung

Bei der Entwicklung verschiedener Prognosemodelle – ob lineare, logistische oder auch Überlebenszeitmodelle - wird man vor zwei Herausforderungen gestellt: die Trennung der wesentlichen Einflussvariablen von weiteren sogenannten Störvariablen und die Handhabung fehlender Werte. Viele Methoden zur Variablenselektion, wie z.B. Stepwise-Selektion, und deren Eigenschaften sind bekannt und die Umsetzung ist in SAS implementiert. In der Regel werden diese Verfahren jedoch nur im Rahmen einer Complete Case Analyse durchgeführt, bei welcher nur die Beobachtungen mit vollständigem Wertevektor zur Berechnung verwendet werden. Von dieser Vorgehensweise ist jedoch bekannt, dass sie verzerrte Ergebnisse zur Folge haben können, da die fehlenden Werte selten rein zufällig vorkommen. Der Ansatz der multiplen Imputation versucht sowohl dieser Verzerrung vorzubeugen und die Power der Analyse zu erhöhen, während der Unsicherheit der Ersetzung Rechnung getragen wird. Derzeit finden sich leider nur wenige Literaturstellen, wie die Variablenselektion auf multipel imputierte Datensätze angewandt werden kann. Im Rahmen einer Analyse basierend auf simulierten Datensätzen für eine logistische Regression wurde nun untersucht, wie verschiedene Methoden zum Umgang mit fehlenden Werten sowie zur Variablenselektion kombiniert werden können und welche Auswirkung dies auf die Modellzusammensetzung hat. Dabei wurden im Wesentlichen Bootstrap-Verfahren berücksichtigt. In dem Beitrag stellen wir die Vorgehensweise, die Umsetzung in SAS sowie die Ergebnisse der Simulationsstudie vor.

Schlüsselwörter: fehlende Werte, Imputation, PROC MI, Variablenselektion, Prognosemodelle, PROC LOGISTIC

1 Einleitung

Bei der Entwicklung verschiedener Prognosemodelle – ob lineare, logistische oder auch Überlebenszeitmodelle - wird man vor zwei Herausforderungen gestellt: die Trennung der wesentlichen Einflussvariablen von weiteren sogenannten Störvariablen und die Handhabung fehlender Werte. Viele Methoden zur Variablenselektion und deren Eigenschaften sind bekannt [1] und die Umsetzung ist in SAS implementiert. In der Regel werden diese Verfahren jedoch nur im Rahmen einer Complete Case Analyse durchgeführt, bei welcher nur die Beobachtungen mit vollständigem Wertevektor zur Berechnung verwendet werden. Von dieser Vorgehensweise ist jedoch bekannt, dass sie verzerrte Ergebnisse zur Folge haben können, da die fehlenden Werte selten rein zufällig vorkommen. Der Ansatz der multiplen Imputation versucht sowohl dieser Verzerrung vorzubeugen und die Power der Analyse zu erhöhen, während der Unsicherheit der Ersetzung Rechnung getragen wird. Derzeit finden sich leider nur wenige Literaturstellen, wie die Variablenselektion auf multipel imputierte Datensätze angewandt werden kann [2].
Im Rahmen einer umfangreichen Analyse [3] basierend auf simulierten Datensätzen für eine logistische Regression [4] wurde nun untersucht, wie verschiedene Methoden zum Umgang mit fehlenden Werten sowie zur Variablenselektion kombiniert werden können und welche Auswirkung dies auf die Modellzusammensetzung hat. Dabei wurden im Wesentlichen Bootstrap-Verfahren berücksichtigt.

2 Material und Methodik

2.1 Daten

Die simulierten Datensätze basieren auf 30 Kovariaten, welche zu sechs verschiedenen Clustern mit unterschiedlichen Korrelationen zwischen den Variablen gehören, um den Einfluss von Multikollinearitäten mit untersuchen zu können und nahe an realen Datensituationen in großen Projekten zur Entwicklung von Prognosemodellen zu sein. Es gab zwei Cluster ohne Korrelation, zwei Cluster mit einer moderaten Korrelation und zwei Cluster mit einer hohen Korrelation zwischen den jeweils fünf Kovariaten. Nur die erste und zweite Variable jedes Clusters war mit dem Zielwert assoziiert. Diese wahren Prädiktoren in den ersten drei Clustern hatten einen Regressionskoeffizienten von 0.8 und diejenigen in den letzten drei Clustern hatten einen Regressionskoeffizienten von 0.4 (s. Tabelle 1).
Verschiedene Typen von Kovariaten, zugrundeliegende Datenmechanismen der fehlenden Werte, Fallzahlen und Fall-Kontroll-Verhältnisse wurden kombiniert und resultierten in 32 verschiedenen simulierten Datensätzen (s. Tabelle 2). Als Referenzszenario wurde das Szenario mit einer Fallzahl von 500, einem Fall-Kontroll-Verhältnis von 1:1, ausschließlich standardnormalverteilten Variablen mit nur wenigen fehlenden Werten eines vollständig zufälligen Datenmechanismus (MCAR) gewählt.

Tabelle 1: Informationen zu den simulierten Kovariaten

Cluster	**Variablen**	**Korrelation innerhalb des Clusters**	**Regressionskoeffizienten**
Cluster a	x_1 - x_5	$\rho \sim 0.0$	$\beta_{1\text{-}2} = 0.8$, $\beta_{3\text{-}5} = 0$
Cluster b	x_6 - x_{10}	$\rho \sim 0.4$	$\beta_{6\text{-}7} = 0.8$, $\beta_{8\text{-}10} = 0$
Cluster c	x_{11} - x_{15}	$\rho \sim 0.9$	$\beta_{11\text{-}12} = 0.8$, $\beta_{13\text{-}15} = 0$
Cluster d	x_{16} - x_{20}	$\rho \sim 0.0$	$\beta_{16\text{-}17} = 0.4$, $\beta_{18\text{-}20} = 0$
Cluster e	x_{21} - x_{25}	$\rho \sim 0.4$	$\beta_{21\text{-}22} = 0.4$, $\beta_{23\text{-}25} = 0$
Cluster f	x_{26} - x_{30}	$\rho \sim 0.9$	$\beta_{26\text{-}27} = 0.4$, $\beta_{28\text{-}30} = 0$

Tabelle 2: Informationen zu den simulierten Szenarien

Szenario Charakteristik	**Ausprägungen**	
Fallzahl	500, 5000	2 Ausprägungen
Verteilung der Kovariaten	standardnormalverteilt, binär	2 Ausprägungen
Fall-Kontroll-Verhältnisse	1:1, 1:3	2 Ausprägungen
Datenmechanismen der fehlenden Werte	MCAR ($0\% \leq p[miss] \leq 7.5\%$), MCAR ($0\% \leq p[miss] \leq 15\%$), MAR ($0\% \leq p[miss] \leq 10\%$), MAR ($0\% \leq p[miss] \leq 20\%$)	4 Ausprägungen
		32 Kombinationen

2.2 Analysenprozedur

In jedem simulierten Datensatz wurden 12 verschieden Analyseverfahren durchgeführt, welche auf drei unterschiedlichen Methoden zum Umgang mit fehlenden Werten und auf vier unterschiedlichen Variablenselektionsverfahren beruhten.

Tabelle 3: Informationen über die zugrundeliegenden statistischen Methoden

Umgang mit fehlenden Werten	**Variablenselektion**
CC: Complete Case (keine Imputation) EM: EM Algorithmus (einfache Imputation) FCS: FCS Algorithmus (fünffache Imputation)	S1: Rückwärtselimination (p=0.157) S2: schrittweise Elimination (p=0.157, p=0.150) S3: LASSO penalisierte Regression (AIC) S4: Bootstrap Selektion (100 Samples, Inklusionsrate ≥ 80%)

Die Spezifikationen für die einzelnen Methoden sind in Klammern angegeben und beruhen auf Hinweisen aus der Literatur (FCS: [5], S1+S2: [6-10], S4: [11-13]). Eine besondere Schwierigkeit stellte die Kombination der fünffachen Imputation mit den Variablenselektionsverfahren dar. Im Rahmen dieser Analyse wurde hierfür ein Verfahren angewandt, welches dem Bootstrapselektionsverfahren ähnelt [14]:

- Schritt 1: Generierung von fünf imputierten Datensätzen (5 FCS) aus dem Originaldatensatz (OS)
- Schritt 2: Variablenselektion in jedem der imputierten Datensätze und Identifikation von Variablensets
- Schritt 3: Identifikation eines finalen Sets an Variablen, das nur Variablen enthält, welche in mindestens 80% der imputierten Datensätze ausgewählt wurden
- Schritt 4: Schätzung der Regressionskoeffizienten des logistischen Regressionsmodells für jeden imputierten Datensatz
- Schritt 5: Mittelung der Regressionskoeffizienten entsprechend Rubin‘s Rule

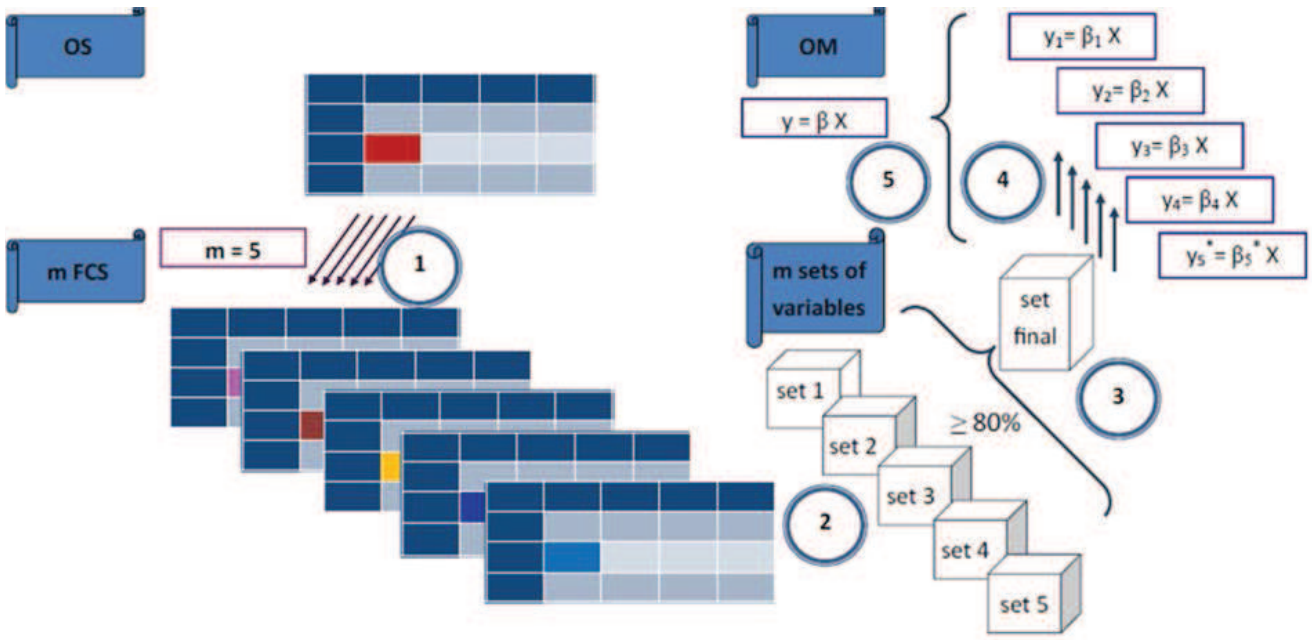

Abbildung 1: Visualisierung der Vorgehensweise

2.3 Zusammenstellung der Ergebnisse

In jedem der 32 simulierten Datensätze wurden diese beschriebenen 12 verschiedenen Analyseverfahren durchgeführt und es wurde jeweils ein finales Set an Prädiktoren identifiziert. Für diese Variablen wurden dann auch die Regressionskoeffizienten des logistischen Regressionsmodells geschätzt. Die Genauigkeit des Models (gemessen am c-Index) wurde geschätzt und mit Bootstrap Samples validiert [15-16]. Die einzelnen untersuchten Kenngrößen sind in Tabelle 4 dargestellt.

Tabelle 1: Informationen über die untersuchten Kenngrößen

Bereich	**Kenngröße**
Genauigkeit	unkorrigierter c-Index
	C-Index korrigiert nach Enhanced Bootstrap Validation
	Optimismus entsprechend der Enhanced Bootstrap Validation
Variablenset	Anzahl der ausgewählten Variablen (size)
	True Positive Rate (TP) = Anteil der wahren ausgewählten Prädiktoren (β>0) an allen ausgewählten Variablen
	Identifikationsrate (ID) = Anteil der wahren ausgewählten Prädiktoren (β>0) an allen wahren Prädiktoren IDs wurden für folgende Subgruppen an Prädiktoren ermittelt: o β>0: alle wahren Prädiktoren o β=0.4: schwache wahre Prädiktoren o β=0.8: starke wahre Prädiktoren o β>0, ρ~0: alle wahren Prädiktoren aus Clustern ohne Korrelation zwischen den Variablen o β>0, ρ~0.4: alle wahren Prädiktoren aus Clustern mit moderater Korrelation zwischen den Variablen o β>0, ρ~0.9: alle wahren Prädiktoren aus Clustern mit hoher Korrelation zwischen den Variablen

Für jede Kenngröße ergaben sich 384 Beobachtungen basierend auf den 12 Analyseprozeduren, welche wiederum auf 32 simulierte Datensätze angewandt worden sind. Um das Ergebnis zusammenzufassen, wurde für jede der Kenngrößen aus Tabelle 4 ein allgemeines lineares Modell mit Hilfe der GLM Prozedur in SAS angepasst. Diese Analyse erhebt keinen Anspruch einer konfirmatorischen Auswertung sondern war ausschließlich dazu gedacht die Vielfalt der Ergebnisse auf deskriptive Weise zusammenzufassen. Das allgemeine lineare Modell umfasste als unabhängige Variablen, die Analysenprozedur und die Szenariocharakteristiken (siehe Formel 1).

Formel 1: Deskriptiver Ansatz um mit Hilfe eines allgemeinen linearen Modells die Ergebnisse zusammenzufassen

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

x_1: Methode zum Umgang mit fehlenden Werten
x_2: Methode zur Variablenselektion
$x_1 * x_2$: Interaktion zwischen Methode zum Umgang mit fehlenden Werten und Methode zur Variablenselektion
x_4: Fallzahl
x_5: Art der Kovariaten
x_6: Fall-Kontroll-Verhältnis
x_7: Datenmechanismus der fehlenden Werte

2.4 Umsetzung in SAS

Im Folgenden wird ausschließlich, um die Ausführung kurz zu halten, die Umsetzung in SAS der Kombination der fünffachen Imputation mit den Variablenselektionsverfahren am Beispiel der Rückwärtselimination dargestellt. Die verwendeten Makrovariablen und die Namen der Datensätze sind in Abbildung 2 zusammengestellt.

OSdata	Originaldatensatz
&allvar	Makrovariable aus allen (binären) Variablen %let allvar= x1-x30;
FCSdata	imputierte Datensätze
&nimp	Makrovariable über die Anzahl an Imputationen %let nimp=5;
FCSS1modvar	Hilfsdatensatz
FCSS1modfreq	Hilfsdatensatz
FCSS1modres	Hilfsdatensatz
FCSS1set	finaler Datensatz mit finalem Variablenset
&inclfreq	Makrovariable über die Grenze bzgl. der Einschlussfrequenz %let inclfreq=0.8;

Abbildung 2: Übersicht verwendeter Datensatzbezeichnungen und Makrovariablen

2.4.1 Multiple Imputation in SAS

Für die Imputation wurde die MI Prozedur in SAS verwendet. Für kontinuierliche Variablen kann eine Spannweite und die Einheit der imputierten Werte spezifiziert werden. Die Spannweite beruhte auf dem minimalen und dem maximalen Beobachtungswert.
Innerhalb der MI Prozedur sind verschiedene Methoden des FCS Algorithmus verfügbar. Die Regressionsmethode wurde ausgewählt für stetige Variablen (statt dem prädiktiven Mittelwert Matching) und die logistische Regressionsmethode für kategoriale Variablen (statt der Diskriminantenmethode). Die logistische Regressionsmethode hat den Vorteil, dass alle explanatorischen Variablen für die Imputation genutzt werden können, während die diskriminante Methode ausschließlich stetige erklärende Variablen verwendet. Zehn Burn-In Iterationen wurden als ausreichend angesehen um zufriedenstellende Ergebnisse zu erzielen [17].

```
proc mi data=OSdata seed=54321
        nimpute= &nimp out=FCSdata;
        class &allvar;
        fcs nbiter=10 logistic(&allvar/details);
        var &allvar;
run;
```

2.4.2 Variablenselektion in SAS

Die sequenziellen Selektionsstrategien, Rückwärtselimination und schrittweise Selektion, wurden innerhalb der LOGISTIC Prozedur in SAS umgesetzt. Die Bootstrapselektion wurde ebenfalls basierend auf der Rückwärtsselektion und somit im Rahmen der LOGISTIC Prozedur in SAS umgesetzt. Die notwendigen Hilfsmakros werden an dieser Stelle jedoch nicht präsentiert. Die LASSO penalisierte Regression wurde mittels der

GLMSELECT Prozedur in SAS umgesetzt, obwohl diese eigentlich für lineare Regressionsmodelle und nicht für logistische Regressionsmodelle ausgelegt ist. Für die Rückwärtselimination wurde der Schwellenwert auf p=0.157 festgelegt, da dieser ähnliche Ergebnisse wie die All-Subset-Selektion mit AIC Kriterion [18] gezeigt hat. Über das BY statement der LOGISTIC Prozedur konnte programmiert werden, dass die Rückwärtselimination in jedem imputierten Datensatz separat erfolgen soll.

```
ods select 'Parameter Estimates';
ods output 'Parameter Estimates'=FCSS1modvar;
proc logistic data=FCSdata;
              class &allvar;
              model outcome (event='1') = &allvar /
              selection backward slstay=0.157 lackfit;
              by _imputation_;
run;
```

2.4.3 Identifikation des finalen Sets in SAS

Zur Identifikation des finalen Sets an Variablen in mindestens 80% der imputierten Datensätze wurde eine Häufigkeitstabelle mit Hilfe der FREQ Prozedur auf den „Parameter Estimates“ Output der LOGISTIC Prozedur angewandt. In einem weiteren Datenschritt wurde dann die Inklusionsfrequenz berechnet indem die Häufigkeit aus der FREQ Prozedur durch die Anzahl der imputierten Datensätze (als Makrovariable hinterlegt) geteilt wurde. Die Zeile zum „Intercept“ wurde gelöscht. Zuletzt wurden dann alle Beobachtungen gelöscht für welche die Inklusionsfrequenz den Schwellenwert (ebenfalls als Makrovariable hinterlegt) nicht erreichte.

```
ods select 'one-way frequencies';
ods output 'one-way frequencies'=FCSS1modfreq;
proc freq data= FCSS1modvar;
     tables variable;
run;

data FCSS1modres (keep=variable inclfreq);
     set FCSS1modfreq;
     inclfreq=frequency/&nimp;
     if variable="Intercept" then delete;
run;

data FCSS1set;
     set FCSS1modres;
     where inclfreq ge &inclfreq;
run;
```

2.4.4 Schätzung der Regressionskoeffizienten in SAS

Aus dieser Datei (FCSS1set) konnte dann in verschiedenen Datenschritten und unter Verwendung der SQL Prozedur und der TRANSPOSE Prozedur das finale Set an Prädiktoren als Makrovariable *[&FCSS1modset]* ausgegeben werden. Die Regressionskoeffizienten für diese Variablen wurden dann mittels LOGISTIC Prozedur *[model*

outcome (event='1') = &FCSS1modset; by _imputation_;] geschätzt und der Mittelwert zu den jeweiligen Schätzern wurde berechnet.

3 Ergebnisse und Diskussion

Im Rahmen der Analyse basierend auf simulierten Datensätzen wurde untersucht, wie verschiedene Methoden zum Umgang mit fehlenden Werten sowie zur Variablenselektion kombiniert werden können und welche Auswirkung dies auf die Modellzusammensetzung hat. Dabei wurden im Wesentlichen Bootstrap-Verfahren berücksichtigt.

3.1 Anwendbarkeit

Die verschiedenen Methoden zum Umgang mit fehlenden Werten sowie zur Variablenselektion konnten miteinander kombiniert werden und in SAS zufriedenstellend umgesetzt werden. Es hat sich jedoch gezeigt, dass bereits in der Umsetzung der einzelnen Methoden verschiedene Schwierigkeiten auftreten. Insbesondere für unerfahrene Statistiker, stellt die Auswahl der geeigneten Methoden (z.B. verschiedene Variablenselektionsmethoden, verschiedene Methoden innerhalb des FCS Algorithmus, …) und der Schwellenwerte (z.B. Anzahl der Imputationen, Anzahl der Bootstrapsamples, p-Werte, Einschlussfrequenz, …) eine große Herausforderung dar, welche das Gesamtergebnis deutlich beeinflussen kann.
Ein weiteres Problem ist, dass der EM Algorithmus eine multivariate Normalverteilung der Daten voraussetzt. Im Rahmen dieser Arbeit wurde jedoch dem Hinweis gefolgt, dass für binäre und kategorielle Variablen die Werte auf den nächstmöglichen Wert gerundet werden könnten [19].

Aber es gibt auch Probleme bei Methoden, welche in SAS noch nicht ausreichend implementiert sind. So wurde die LASSO penalisierte Regression in SAS bisher nur für lineare Modelle implementiert. Die GLMSELECT Prozedur beruht auf der Kleinstquadrateschätzung und ist deswegen theoretisch nicht für eine logistische Regression geeignet. Dennoch wurde diese Methode im Rahmen dieser Arbeit verwendet, da es bereits verschiedene Hinweise gab, dass die GLMSELECT Prozedur dennoch dafür verwendet werden könne [20-21].

3.2 Einfluss der verschiedenen Analyseprozeduren und Szenarien

Es hat sich gezeigt, dass ein Modell kaum mehrmals identifiziert wurde. Die Modellzusammensetzung hängt ab von den verwendeten Analyseprozeduren und den Szenariencharakteristiken. Es konnte im Rahmen der Analyse gezeigt werden, dass viele Modelle mit unterschiedlichen Kovariaten eine vergleichbare Modellgenauigkeit aufweisen. Jedoch bei näherer Betrachtung konnte man grundlegende Unterschiede zwischen den verschiedenen Analyseprozeduren erkennen. Die Methoden unterschieden sich bezüglich der Auswahl wahrer Prädiktoren, dem Ausschluss von Störvariablen und der Genauigkeit der Schätzung von Regressionskoeffizienten.

Im Rahmen der Analyse simulierter Daten konnte kein eindeutiger Zusatznutzen der multiplen Imputation mittels FCS-Algorithmus gegenüber der einfachen Imputation mittels EM-Algorithmus dargestellt werden. Dies kann jedoch auch darauf zurückgeführt werden, dass im Rahmen der bisherigen Analysen die zusätzliche Variabilität bei der Imputation fehlender Werte, welche sich in der Varianz widerspiegelt, nicht berücksichtigt wurde. Im Rahmen der multiplen Imputation wurde ein gesondertes Selektionsverfahren durchgeführt, welches auf dem Bootstrap-Selektionsverfahren beruht. Dieses angepasste Verfahren zeigte wie das Boostrap-Selektionsverfahren selbst eine höhere Effizienz beim Ausschluss von Störvariablen.

Eine detaillierte Durchführungsbeschreibung und Dokumentation der Ergebnisse ist in [3] beschrieben. Die Arbeit kann bei Interesse bei der Autorin per Mail angefordert werden.

4 Ausblick

In weiteren Analysen soll nun der Prozess noch optimiert werden, indem zum Beispiel eine größere Anzahl an Datensätzen über multiple Imputation ($m > 5$) generiert wird. Außerdem kann der Grenzwert, welcher für die bisherige Analyse auf 80% gesetzt wurde, noch abgesenkt werden (< 80%) um weniger restriktiv zu sein und somit weniger wahre Prädiktoren fälschlicherweise auszuschließen. Eine weitere Möglichkeit wäre, nicht nur die Regressionskoeffizienten selbst sondern auch deren Varianz entsprechend Rubin‘s Rule zu schätzen.

Literatur

[1] Steyerberg EW: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: Springer-Verlag; 2009

[2] Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet, Henrica C W: Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol 2007, 7:33

[3] Stierlin AS: Assessment of stability and validity of logistic regression models using bootstrap resampling in simulated data as well as in real data: To what extent does the approach of missing data imputation and variable selection affect the composition and performance of logistic regression models? Master Thesis. Ruprecht-Karls-Universität Heidelberg; 2014

[4] Wicklin R: Simulating data with SAS. Cary, N.C: SAS Institute; 2013

[5] Rubin DB: Multiple Imputation after 18+ years. Journal of the American Statistical Association 1996, 91:473-489.

[6] Lee K in, Koval JJ: Determination of the best significance level in forward stepwise logistic regression. Communications in Statistics - Simulation and Computation 1997, 26:559-575.

[7] Hosmer DW, Lemeshow S: Applied logistic regression. 2nd edition. New York: Wiley; 2000 [Wiley series in probability and statistics. Texts and references section].

[8] Hocking RR: Methods and applications of linear models: Regression and the analysis of variance; 2013 [Wiley series in probability and statistics].

[9] Stoddard GJ: Biostatistics and epidemiology using STATA: A Course Manual. University of Utah; 2010.

[10] Teräsvirta T, Mellin I: Model Selection Criteria and Model Selection Tests in Regression Models. Scandinavian Journal of Statistics 1986, 13:159-171.

[11] Austin PC, Tu JV: Bootstrap Methods for Developing Predictive Models. The American Statistician 2004, 58:131-137.

[12] Sauerbrei W, Schumacher M: A bootstrap resampling procedure for model building: Application to the cox regression model. Statist. Med. 1992, 11:2093-2109.

[13] Sauerbrei W: The use of resampling methods to simplify regression models in medical statistics. Journal of Applied Statistics 1999, 48:313-329.

[14] Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet, Henrica C W: Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol 2007, 7:33.

[15] Muche R, Ring C, Ziegler C: Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Aachen: Shaker Verlag; 2005

[16] Muche R: Validierug von Regressionsmodellen: Notwendigkeit und Beschreibung der wichtigsten Methoden. Die Rehabilitation 2008, 47: 56-62

[17] SAS support: The MI procedure. FCS Methods for Data Sets with Arbitrary Missing Patterns [http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect026.htm].

[18] Teräsvirta T, Mellin I: Model Selection Criteria and Model Selection Tests in Regression Models. Scandinavian Journal of Statistics 1986, 13:159-171.

[19] Baneshi MR, Talei AR: Does the missing data imputation method affect the composition and performance of prognostic models? Iran Red Crescent Med J 2012, 14:31-36.

[20] Flom PL, Cassell DL: Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In Proceedings of NESUG conference 2007. Edited by Williams C, Mitchell R; 2007.

[21] Flom PL: Multinomial and ordinal logistic regression using PROC LOGISTIC. In NESUG 2010. The Proceeding of the 23rd annual conference of the NorthEast SAS Users Group. Edited by NESUG; 2010.