Die Anpassung eines gewöhnlichen logistischen Regressionsmodells für alle möglichen Kombinationen von Datensätzen einer Beobachtungsstudie – Ein Beispiel für den Umgang mit rechenintensiven Prozessen und großen Dateien

Christina Ring*,
Rainer Muche*,
*Institut für Biometrie,
Universität Ulm
Schwabstr. 13
89075 Ulm
christina.ring@uni-ulm.de

Jens Dreyhaupt*,
Siegfried Wieshammer#
#Pneum.-Thoraxch. Zentrum,
Klinikum Offenburg
Ebertplatz 12
77654 Offenburg

Zusammenfassung

In der klinischen Forschung treten häufig komplexe Datensätze auf, bei denen pro Untersuchungseinheit für jedes Merkmal mehr als nur eine Beobachtung erhoben wird. Da wegen der Abhängigkeit von Beobachtungen die Anwendung vieler „klassischer" statistischer Verfahren nicht zulässig ist, kann in einer solchen Situation – neben der Anwendung adäquater statistischer Methoden für abhängige Beobachtungen – die Durchführung einer Simulationsstudie basierend auf Datensätzen mit unabhängigen Beobachtungen erfolgen. Die Erfahrungen und Ergebnisse der Anwendung des letztgenannten Verfahrens sind Inhalt dieses Artikels.

Grundlage der Arbeit sind die Daten einer nichtrandomisierten Beobachtungsstudie an 223 Ambulanzpatienten mit Atemnot, die insgesamt 248 Inhalatoren benutzen. Ein wichtiges Ziel bei der Auswertung der Studie bestand in der Untersuchung des populationsbezogenen Einflusses (marginales Modell) verschiedener potentieller prognostischer Faktoren (Lebensalter, Art des Inhalators, Schweregrad der Obstruktion und Schulung im Umgang mit dem Inhalator) auf das Auftreten einer ineffektiven Inhalation. Wegen der Mehrfachbeobachtungen pro Untersuchungseinheit (199 Patienten verwendeten einen Inhalator, 23 Patienten nutzten zwei Inhalatoren, ein Patient nutzte 3 Inhalatoren), ist die Anwendung eines gewöhnlichen logistischen Regressionsmodells nicht angemessen. Eine Möglichkeit, das marginale Modell zu erzeugen, besteht in der Erstellung aller möglichen Kombinationen, bei welchen von jedem der 24 Patienten, die mehr als einen Inhalator benutzten, genau ein Inhalator für die Analyse ausgewählt wird (Erzeugung von Datensätzen mit unabhängigen Beobachtungen). Für die gegebene Situation erhält man auf diese Art $2^{23} \cdot 3 = 25.165.824$ verschiedene Datensätze mit unabhängigen Beobachtungen.

Im Artikel wird das Vorgehen der Anpassung eines gewöhnlichen logistischen Regressionsmodells für die 25.165.824 verschiedenen Datensätze vorgestellt und mögliche Lösungen für die Arbeit mit rechenintensiven Prozessen und großen Dateien angegeben. Die Arbeiten wurden am Institut für Biometrie der Universität Ulm unter Verwendung von SAS® 9.1 und 9.2 auf 14 Windows PCs durchgeführt. Sie waren Teil eines Methodenvergleichs.

Schlüsselwörter: Marginales Modell, logistische Regression, Erzeugung von Datensätzen mit unabhängigen Beobachtungen, Simulationsstudie

1 Einleitung und Fragestellung

In der klinischen Forschung treten häufig komplexe Datensätze auf, bei denen pro Untersuchungseinheit für jedes Merkmal mehr als nur eine Beobachtung erhoben wird. Beispiele hierfür sind Messungen über den zeitlichen Verlauf von Merkmalen, Messungen ein- und desselben Merkmals an verschiedenen Stellen des Körpers von Patienten, Messungen ein- und desselben Merkmals an paarigen Organen. In derartigen Datensätzen treten teilweise komplexe Abhängigkeitsstrukturen auf, so dass die Anwendung vieler klassischer statistischer Verfahren nicht zulässig ist. In einer solchen Situation kann, neben der Anwendung adäquater statistischer Methoden für abhängige Beobachtungen, die Durchführung einer Simulationsstudie basierend auf Datensätzen mit unabhängigen Beobachtungen erfolgen. Die Erfahrungen und Ergebnisse der Anwendung des letztgenannten Verfahrens sind Inhalt dieses Artikels.

Medizinischer Hintergrund

Pulverinhalatoren werden häufig fehlerhaft angewendet, was zu einer ineffektiven Inhalation führt. Zur Untersuchung dieser Fragestellung wurde am Pneumologisch-Thoraxchirurgischen Zentrum des Klinikums Offenburg eine nichtrandomisierte Beobachtungsstudie an 223 Ambulanzpatienten mit Atemnot durchgeführt, die insgesamt 248 Inhalatoren benutzen. Das waren Patienten mit Asthma, COPD oder anderen Lungenerkrankungen. Neben der Erhebung anamnestischer Daten erfolgten körperliche sowie umfangreiche apparative Untersuchungen. Zur Beurteilung der Effektivität der Inhalation mussten die Patienten für jeden von ihnen genutzten Inhalatortyp einen Inhalationsvorgang demonstrieren. Im Ergebnis einer vorhergehenden Arbeit wurden Lebensalter, Art des Inhalators, Schweregrad der Obstruktion und Schulung im Umgang mit dem Inhalator als Faktoren mit wesentlichem Einfluss auf das Auftreten einer ineffektiven Inhalation identifiziert [1].

Das Ziel dieser Arbeit bestand in der Quantifizierung der Effekte der genannten Einflussgrößen auf die Zielgröße „Auftreten einer ineffektiven Inhalation“ (ja/nein) auf Populationsebene (d.h. Angabe populationsbezogener Odds Ratios). Bei ausschließlich unabhängigen Beobachtungen und binärer Zielgröße wäre es möglich, ein logistisches Regressionsmodell anzupassen. Wegen der Mehrfachbeobachtungen (Patienten nutzten zum Teil verschiedene Inhalatoren) kann die Anpassung eines logistischen Regressionsmodells jedoch zu verzerrten Resultaten führen. Stattdessen sollten adäquate statistische Verfahren verwendet werden.

Im Rahmen eines Methodenvergleichs wurden die genannten Daten mit drei verschiedenen statistischen Methoden ausgewertet und die Ergebnisse der (marginalen) Modelle miteinander verglichen [2]. Bei den drei Methoden handelte es sich um ein GEE Modell [3], die Marginalisierung eines gemischten Regressionsmodells und die sogenannte Zensusmethode, deren technische Durchführung und Ergebnisse im Folgenden vorgestellt werden.

2 Methodik

Als Zensusmethode bezeichnen wir die Anpassung jeweils eines gewöhnlichen logistischen Regressionsmodells an alle theoretisch möglichen Datensätze mit unabhängigen Beobachtungen, die sich aus den gegebenen Studiendaten erzeugen lassen. Die Modellergebnisse werden zusammengeführt und anschließend deskriptiv ausgewertet.

Nach einer kurzen Einführung in die Struktur des Studiendatensatzes werden im Folgenden die Datensatzerzeugung, die Anpassung des logistischen Regressionsmodells, die Ergebniszusammenführung sowie die Deskription der Ergebnisse in den wichtigsten Details der programmtechnischen Umsetzung beschrieben.

2.1 Struktur der Studiendaten

Die zur Verfügung stehenden Daten stammen von 223 Patienten, die insgesamt 248 Inhalatoren verwendeten: 199 dieser Patienten nutzten einen Inhalator, 23 Patienten zwei Inhalatoren, ein Patient drei Inhalatoren. Tabelle 1 gibt einen Überblick über die Struktur des Datensatzes; die Variable **patnr** in Spalte 1 beschreibt die eindeutige Patientennummer.

Tabelle 1: Struktur der gegebenen Studiendaten

patnr	**eff**	**alter**	**ger_1**	**obstruktion**	**schulung1**	**…**
...	...	...	...	...	...	...
25	0	73.1	1	0	1	...
26	0	79.0	1	1	0	...
26	1	79.0	2	1	0	...
27	1	71.9	2	1	1	...
...	...	...	...	...	...	...
223	1	20.6	0	2	1	...

Neben der Zielgröße „Auftreten einer ineffektiven Inhalation“ (**eff**, 0 = ineffektive Inhalation, 1 = effektive Inhalation) sind in Tabelle 1 wichtige prognostische Faktoren dargestellt:

- Lebensalter (**alter)** in Jahren
- Art des Inhalators (**ger_1**): 0 = Turbuhaler®, 1 = HandiHaler®, 2 = Discus®, 3 = Aerolizer®
- Schweregrad der Obstruktion (**obstruktion**): 0 = keine Obstruktion, 1 = leichtgradige Obstruktion, 2 = mittelgradige Obstruktion, 3 = schwergradige Obstruktion
- Schulung im Umgang mit dem Inhalator (**schulung1**): 0 = keine Schulung, 1 = Schulung ist erfolgt.

Die Variablen Lebensalter und Schweregrad der Obstruktion sind dabei auf den Patienten bezogen. Wohingegen die Variablen Art des Inhalators und Schulung den jeweiligen Inhalator des betrachteten Patienten betreffen.

Nach der Anzahl der verwendeten Inhalatoren kann der Studiendatensatz in drei Teildatensätze gesplittet werden:

- *Teildatensatz 1*: besteht aus 199 Zeilen der Patienten, die nur einen Inhalator einbringen
- *Teildatensatz 2*: besteht aus 46 Zeilen der 23 Patienten, die zwei Inhalatoren-einbringen (→ 2^{23} = 8 388 608 mögliche Kombinationen)
- *Teildatensatz 3*: besteht aus drei Zeilen: eine Zeile für jeden Inhalator, den der Patient mit den drei Inhalatoren einbringt.

Aus diesen 3 Teildatensätzen werden im Folgenden alle für die Zensusmethode notwendigen und theoretisch möglichen Datensätze mit jeweils 223 unabhängigen Beobachtungen erzeugt. Insgesamt sind das 2^{23}*3 = 25 165 824 Datensätze.

2.2 Datensatzerzeugung

2.2.1 Prinzip

Ausgehend vom Studiendatensatz werden die 2^{23}*3 Datensätze mit unabhängigen Beobachtungen generiert, indem zuerst aus Teildatensatz 2 alle theoretisch möglichen 2^{23} Datensätze mit 23 unabhängigen Beobachtungen erzeugt werden. Diese 8 388 608 Datensätze werden anschließend jeweils um den Teildatensatz 1 ergänzt. Im nächsten und letzten Schritt wird jede der drei Zeilen des Teildatensatzes 3 mit jedem der zuvor erzeugten Datensätze kombiniert, so dass die 2^{23}*3 Datensätze mit jeweils 223 unabhängigen Beobachtungen entstehen. Abbildung 1 veranschaulicht das Vorgehen.

In der programmtechnischen Umsetzung wurden die aus Teildatensatz 2 zu generierenden Datensätze mit Hilfe von 0/1-*Indikatorvektoren* der Dimension 23 selektiert. Dabei war die i-te Komponente des Indikatorvektors (i = 1, …, 23) dem i-ten Patienten des Teildatensatzes 2 zugeordnet und zeigte an, ob für diesen Patienten die Zeile mit dem ersten Inhalator (i-te Komponente = 0) oder die Zeile mit dem zweiten Inhalator (i-te Komponente = 1) verwendet werden sollte. Ein einfaches „Mergen“ des Indikator-Datasets mit dem Teildatensatz 2 (der eine entsprechende Variable enthielt) führte auf die gewünschte Selektion.

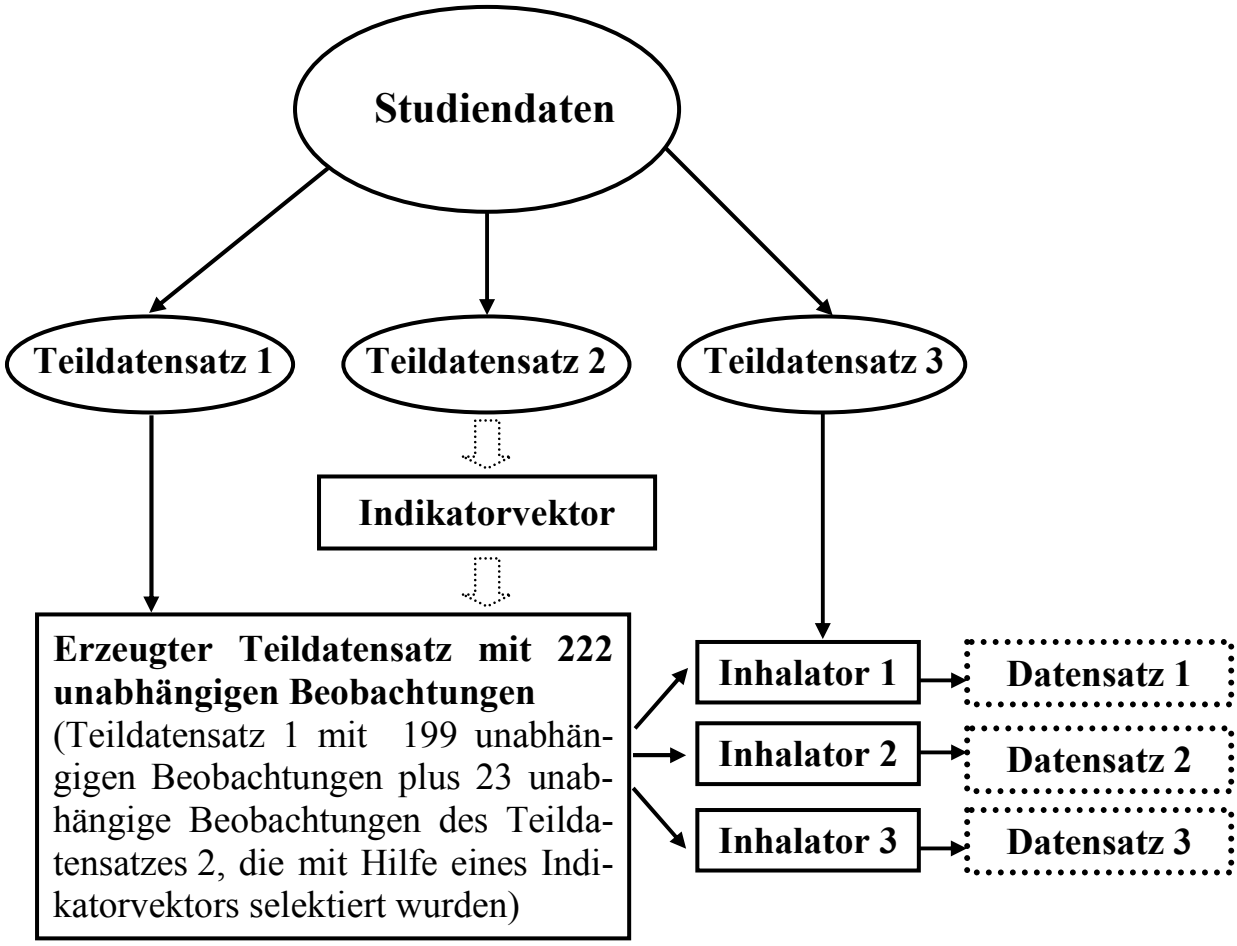

Abbildung 1: Schema der Erzeugung von Datensätzen mit 223 unabhängigen Beobachtungen aus den Teildatensätzen 1, 2 und 3. Die drei Teildatensätze sind disjunkt und ergeben vereinigt den Studiendatensatz. Je Indikatorvektor entstehen 3 der $2^{23} \cdot 3$ theoretisch möglichen Datensätze mit jeweils 223 unabhängigen Beobachtungen (Datensatz 1 bis 3).

2.2.2 Erzeugung der Indikatorvektoren

Zur Erzeugung der Indikatorvektoren wurden die Zahlen 0, 1, …, 2^{23}-1 als Binärzahlen dargestellt; jede dieser Binärzahlen entspricht einem Indikatorvektor. Um das Zusammenspiel der public domain Software *R* und SAS® zu testen, wurde dieser Teil der Programmierung mit *R* realisiert. Der folgende *R*-Code zeigt die Umsetzung:

```
conv.bin<-function(zahl,stellen=23){
  v<-rep(0,stellen) # Nullvektor
  i<-0              # Laufindex Nullinitialisierung
  z<-zahl
  while (z>=1){
    v[stellen-i]<-z%%2   # Rest der Mod 2 Berechnung
    # Ergebnis der Mod 2 Berechung (welches Vielfache von 2):
    z<-z%/%2
    i<-i+1} return(v)}
```

Jeweils 2500 dieser Indikatorvektoren wurden in einer Datei als sogenanntes Paket zusammengefasst. Insgesamt ergaben sich auf Grund der benötigten 2^{23} Indikatorvektoren 3356 solcher Pakete, wobei die letzte Datei nur 1108 Indikatorvektoren enthielt. Diese Aufsplittung war sinnvoll, da die anschließenden Berechnungen in bestimmten Zeitfenstern und auf 14 verschiedenen PCs erfolgen sollten.

Der Export der Pakete von R nach SAS® konnte sehr leicht realisiert werden:

```
library(foreign)
write.foreign(res,datafile=f.txt,codefile=f.sas,package="SAS",
          dataname=paste("Teil",nr.teil,sep=""),validvarname="V7")

# Inhalt der erzeugten Dateien ansehen
file.show(f.txt) # Daten
file.show(f.sas) # SAS Einleseprogramm
```

Im Ergebnis dieses *R*-Codes entstanden jeweils eine Textdatei mit den Indikatorvektoren und eine SAS®-Programmdatei zum Einlesen der Indikatorvektoren aus der Textdatei. Die Namen dieser beiden Dateien wurden so gewählt, dass über eine fortlaufende Nummer das genaue Paket identifiziert werden konnte (siehe Argument `dataname` der Funktion `write.foreign`).

Der folgende SAS®-Quellcode zeigt eines dieser SAS®-Einleseprogramme, die Datei „Teil_0_2009-06-09.sas“, die die ersten 2500 Indikatorvektoren (bezeichnet mit S0 bis S2499) aus der Textdatei „Teil_0_2009-06-09.txt“ nach SAS® einliest.

```
DATA  Teil0 ;
  INFILE  "F:/GMDS_2009/R OUTPUT/Teil_0_2009-06-09.txt"
       DSD
       LRECL= 5003 ;
  INPUT
   S0
   S1
   S2
   …
   S2499;
RUN;
```

2.3 Anpassung des logistischen Regressionsmodells

Simulationsumgebung
Die praktische Realisierung der Modellanpassung erfolgte unter Verwendung der SAS® Versionen 9.1 und 9.2, verteilt auf 14 Windows PCs mit folgenden Konfigurationen:

2 x Pentium 4, 3.0 GHz, 0.5 GB RAM
1 x Pentium 4, 3.4 GHz, 2 GB RAM
7 x Pentium 4, 3.0 GHz, 1 GB RAM
1 x Dual Core, 3.4 GHz, 1 GB RAM
3 x Core TM2, 2.0 GHz, 2 GB RAM

Vor dem Start der Simulationsrechnungen musste auf jedem PC die Simulationsumgebung eingerichtet werden:

- ein Verzeichnis für die Pakete mit den Indikatorvektoren und die zugehörigen SAS®-Einleseprogramme (siehe Abschnitt 2.2.2)
- ein Verzeichnis für die SAS®-Programme der Auswertung und
- ein Verzeichnis für die Ergebnisdateien der Modellanpassung (siehe unten)

Programmstruktur
Um die 25 165 824 Datensätze entsprechend der Indikatorvektoren erzeugen und jeweils das Regressionsmodell anpassen zu können, war die Mithilfe der Kollegen erforderlich. Sie starteten am Ende ihres Arbeitstages die Programme, die in der Nacht die Simulationen und Auswertungen durchführten und paketweise abspeicherten.

Für einen effizienten und möglichst fehlerfreien Arbeitsablauf durch die verschiedenen „Anwender“ wurde die programmtechnische Umsetzung über zwei separate SAS®-Programme realisiert: ein Start- und ein Auswertungsprogramm.

Startprogramm
Das Startprogramm dient dem einfachen Start der Datenerzeugung und -auswertung. Es ruft das eigentliche Auswertungsprogramm über einen %INCLUDE-Befehl zur Abarbeitung auf.
Im Startprogramm müssen vor jedem erneuten Aufruf lediglich zwei Parameter vom Anwender aktualisiert werden: die Nummer des Paketes, mit dem gestartet werden soll und die Anzahl der Pakete, die abzuarbeiten sind. Diese Parameter steuern das Auswertungsprogramm.

Auswertungsprogramm mit logistischer Regression
Das Auswertungsprogramm enthält mehrere Makros, die über DO Schleifen eine modulare Bearbeitung ermöglichen. Es leistet folgende Teilaufgaben:

- Setzen der Pfade des Verzeichnisses der Quelldaten und des Verzeichnisses der Ergebnisdateien
- Einlesen des Studiendatensatzes und des jeweiligen Paketes

- Für jeden Indikatorvektor: Erzeugung der drei Datensätze mit unabhängigen Beobachtungen (siehe Abschnitt 2.2.2), Anpassung des logistischen Regressionsmodells und Speicherung der Ergebnisse der Modellanpassung

Folgender SAS®-Quellcode wurden im Auswertungsprogramm für das Einlesen der Indikatordateien, die Datensatzerzeugung und die Anpassung des logistischen Regressionsmodells verwendet:

```
* Pakete mit Indikatorvektoren werden eingelesen;
%macro einbinden;
  * nr:   Nummer der ersten Paket-Datei;
  * n_indvecdat:  Anzahl Dateien, die eingelesen werden sollen;
  %do z=&nr %to &nr+(&n_indvecdat-1);
    %include "&source.Teil_&z._&sourcedat..sas";
  %end;
%mend;
%einbinden;
```

```
   * Erzeugung der Datensätze;
   * Je Indikatorvektor 3 Datensätze;
data ind;                       * Datensatz mit Indikatorvektoren;
  set &dat;
  * PTNO und No_geraet zum mergen;
  PTNO=_n_;
  * flag zur Selektion des Gerätes entsprechend S&i;
  flag=1;
  rename S&i=No_geraet;         * S&i ist aktueller Indikatorvektor;
  keep S&i PTNO flag;
run;
…
data inh2s;                     * Datensatz gemäß Indikatorvektor;
  merge inh2 ind;               * inh2 ist Teildatensatz 2;
  by PTNO No_geraet;
  if flag=1;
  drop flag PTNO h;
run;
* Hinzufügen: Patienten mit einem Inhalator und jeweils ein Daten-
* satz des Patienten mit 3 Inhalatoren (Teildatensatz 3 = pat167);
%do j=1 %to 3;
    data ddat;
      set inh2s inh1 pat167;   * inh1 ist Teildatensatz 1;
      if patnr=167 then if No_geraet=&j;
    run;
…
%end;
```

```
ods listing close;  …
ods select ConvergenceStatus Type3 OddsRatios ParameterEstimates;

%macro analysis(dat=Teil1,nstart=1,anz=1);
…
 %do i=&nstart %to &nstart+&anz-1;
    DM 'clear log;';          * Log-Fenster wird gelöscht;
    DM 'odsresults' clear;    * Results-Fenster wird gelöscht;
    * Erzeugung der 3 Datensätze;
    * entsprechend des aktuellen Indikatorvektors (siehe oben);
    …
  %do j=1 %to 3;    * Auswertung der Datensätze;
   …
   ods output ConvergenceStatus=conv;   * einzelne Ergebnisdateien;
   ods output Type3=pvalue;
   ods output OddsRatios=OR;
   ods output ParameterEstimates=parms;
   ods show;
   proc logistic DESCENDING data=ddat; * Logistische Regression;
    title "Schleife &i, Datensatz &j von 3";
    class eff patnr ger_1 obstruktion schulung1 /
                                                Param=Ref Ref=last;
    model eff(event='0')=alter ger_1 obstruktion schulung1/
                                                Risklimits expb;
   run;
   quit;
   ods output close;
   * Ergebnisdatei ergänzen;
   …
  %end;
  * permanente Ergebnisdatei je Paket;
  …
 %end;
…
%mend;
```

2.4 Ergebniszusammenführung

Ergebnisdateien je Paket

Die Datengenerierung und Auswertung erfolgte jeweils für eine Paket-Datei. Deshalb wurden in einem ersten Schritt die Ergebnisse der Modellanpassung eines jeden Paketes in einer SAS®-Datei mit der Bezeichnung „Resteil*.sas7bdat“ (Größe etwa 3.7 MB) zusammengefasst. Der * bezeichnete dabei die Nummer des Paketes, also 0, …, 3355. Bis auf die letzte Ergebnisdatei enthalten diese Resteil-Dateien 2500*3 = 7500 Ergebniszeilen.

Die Ergebnisdateien umfassen alle für die Zielstellung der Arbeit relevanten Ergebnisse der Modellanpassungen, wie z.B. die Schätzer der Odds Ratios (OR) mit oberer und unterer Grenze des zugehörigen Konfidenzintervalls aller eingangs erwähnten Einflussgrößen. Außerdem wurden zu Kontrollzwecken die Bezeichnung des Indikatorvektors sowie der Konvergenzstatus der Modellanpassung dokumentiert.

Für die Erzeugung der Ergebnisdateien enthält das Auswertungsprogramm den folgende SAS®-Quelltext (auszugsweise):

```
* Output-Dateien transponieren;
…
proc transpose data=pvalue out=tpvalue
               prefix=p_Type3_;
  id Effect;
run;
…
* Transponierte Dateien mergen;
* Identifikationsvariable dat=S_ij anfügen;
data res;
  merge conv tpvalue tOR tparms;
  by h;
  dat="                                          ";
  dat="S&i-&j";
run;

…
* Permanente Ergebnis-Datei erzeugen;
data out.res&dat;
  set out.res&dat res; * je Datensatz eine Ergebniszeile angefügt;
run;
```

Abbildung 2 zeigt einen Ausschnitt der Datei „Resteil0.sas7bdat“, der die Struktur dieser Ergebnisdateien verdeutlichen soll.

dat	Konvergenz	OR_LCL_ schulung	OR_est_ schulung	OR_UCL_ schulung	...
S0-1	1	3.094	6.315	12.891	...
S0-2	1	3.179	6.543	13.467	...
S0-3	1	3.049	6.209	12.641	...
S1-1	1	3.116	6.364	12.997	...
S1-2	1	3.199	6.587	13.564	...
S1-3	1	3.059	6.229	12.684	...
...	...	...	...	...	...
S2499-1	1	3.244	6.703	13.848	...
S2499-2	1	3.345	6.978	14.558	...
S2499-3	1	3.176	6.537	13.456	...

Abbildung 2: Struktur der Ergebnisdatei „Resteil0.sas7bdat“.

Gesamtergebnisdateien
Die 3356 Ergebnisdateien der Pakete wurden in mehreren Schritten zusammengeführt; der Einfachheit halber jeweils für die ca. 250 Dateien jedes PCs.

Aufgrund der zu erwartenden Dateigröße wurden bereits in diesem Schritt je PC mehrere Ergebnisdateien mit nur wenigen ausgewählten Variablen erzeugt. Diese Dateien enthalten – separat für jede der Einflussgrößen – das Odds Ratio sowie die obere und untere Grenze des zugehörigen Konfidenzintervalls. Insgesamt entstanden so je PC die folgenden 8 Teilgesamtdateien:

- eine Datei für Alter,
- drei Dateien für Art des Inhalators: Vergleiche Aerolizer® vs. Turbuhaler®, Discus® vs. Turbuhaler®, HandiHaler® vs. Aerolizer®
- drei Dateien für Schweregrad der Obstruktion: Vergleiche 0 vs. 3, 1 vs. 3, 2 vs. 3
- eine Datei für Schulung (nein vs. ja)

Jede dieser 8 Dateien war etwa 121 MB groß und führte bei Vereinigung der 14 Dateien aller PCs auf jeweils eine Gesamtdatei von ca. 1.6 Gigabyte, die anschließend deskriptiv ausgewertet wurde. Falls andere Parameter als die bisher betrachteten Odds Ratios mit Konfidenzintervallen von Interesse wären, müssten aus den 3356 Ergebnisdateien entsprechend neue Gesamtdateien erzeugt werden.

Folgender SAS®-Quelltext wurde für das Aneinanderfügen verwendet:

```
%do k=&startnr %to &endnr;
    data results;
      set results inpath.resteil&k(keep=&varis dat);
    run;
  %end;
```

2.5 Deskription der Modellergebnisse

Die Auswertung der Ergebnisdateien erfolgte mit PROC UNIVARIATE. Wegen der großen Ergebnisdateien war die Deskription nur auf den leistungsstärksten Rechnern möglich. Bei zu geringerem Hauptspeicher hätte man z.B. auf PROC MEANS mit QMETHOD=P2 ausweichen müssen, einer Methode, die die Quantile approximativ berechnet. Je nach Größe des verfügbaren Hauptspeichers ist es empfehlenswert, die Deskription für jede Ergebnisdatei einzeln durchzuführen (jeweils Neustart von SAS®).

Die Ergebnisse der Zensusmethode unterscheiden sich von den Resultaten der beiden anderen Methoden „GEE Modell“ und „Marginalisierung des gemischten Regressionsmodells“: für die interessierenden Parameter erhält man die Verteilungen der Parameter und kann zusätzlich zum Median außerdem den Bereich der möglichen Ausprägungen des jeweiligen Parameters angeben. Tabelle 2 zeigt dies beispielhaft für das Odds Ratio der Variablen Schulung. In Abbildung 3 sind ergänzend die Ergebnisse des Methodenvergleichs für diese Variable exemplarisch dargestellt.

Tabelle 2: Ergebnisse der Zensusmethode am Beispiel des Odds Ratios für Schulung: Deskription der Verteilung des Punktschätzers und der Konfidenzgrenzen. KI: Konfidenzintervall

	Schätzer	Untere Grenze des KI	Obere Grenze des KI
Minimum	5.67	2.81	11.42
1. Quartil	6.40	3.14	13.08
Arithm. Mittel	6.70	3.25	13.81
Median	6.65	3.23	13.69
3. Quartil	6.95	3.35	14.42
Maximum	8.71	4.02	18.88
Standardabweichung	0.41	0.16	1.01
Spannweite	3.04	1.21	7.46
Interquartilsabstand	0.54	0.21	1.34

Für die Variable Schulung führen das GEE Modell und die Marginalisierung des gemischten Modells zu sehr ähnlichen Ergebnissen. Das gilt auch für die anderen betrachteten Einflussvariablen. Im Vergleich dazu weicht die Zensusmethode in ihren quantitativen Ergebnissen von denen der anderen beiden Methoden ab. Das gilt für einige Einflussvariablen mehr, für andere weniger.
Während bei der Zensusmethode durch die gegebene Verteilung der Parameter mehr Informationen zur Verfügung stehen, als bei den anderen beiden Methoden, ist es jedoch nicht möglich, die Abhängigkeitsstruktur in den Mehrfachmessungen (Kovarianzstruktur) abzubilden.

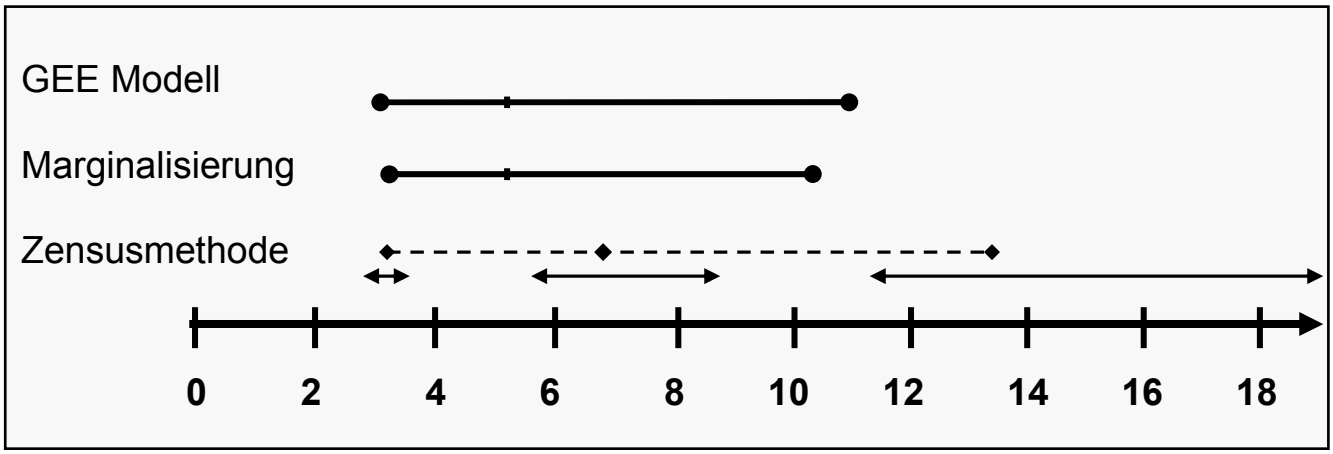

Abbildung 3: Vergleich der Ergebnisse der Anpassung des marginalen Modells für GEE Modell, Marginalisierung des gemischten Regressionsmodells, Zensusmethode am Beispiel des Odds Ratios für Schulung nein vs. ja. Für GEE Modell und Marginalisierung des gemischten Regressionsmodells sind die Schätzer für das OR und die Grenzen des 95% Konfidenzintervalls angegeben. Für die Zensusmethode ist mit der Strichlinie die Verbindung des Medians des Schätzers, des Medians der unteren Grenze und des Medians der oberen Grenze des Konfidenzintervalls angegeben. Die Doppelpfeile geben zusätzlich Lage und Spannweite der Verteilungen von OR, unterer Konfidenzgrenze und oberer Konfidenzgrenze an.

3 Ergebnisse und Diskussion

Die praktische Umsetzung der Zensusmethode erfolgte auf 14 Windows PCs. Für die einzelnen Schritte wurden folgende Laufzeiten ermittelt:

- Erzeugung der 3356 Indikatordateien und SAS®-Einleseprogramme:
  ca. 19 Stunden
- Anpassung des logistischen Regressionsmodells inklusive Einlesen der Indikatordateien, Erzeugung der Datensätze sowie Ergebnisspeicherung in Resteil*.sas7bdat:
  je nach PC zwischen 70 und 120 Minuten pro Paket;
  bei einer durchschnittlich Laufzeit von 90 Minuten je Paket ergeben sich bei 3356 Paketen etwa 7 Monate Rechenzeit, die aufgeteilt auf 14 PCs auf etwa 15 Tage Rechenzeit je PC führen
- Ergebniszusammenführung (Erzeugung der Gesamtdateien):
  je nach PC zwischen 30 bis 40 Minuten für jede Teilgesamtdatei und etwa 50 bis 60 Minuten für die Erzeugung einer Gesamtdatei aus den 14 Teilgesamtdateien; insgesamt sind das etwa 3 Tage
- Deskription der Ergebnisse (Laden der Gesamtdatei in das Work-Verzeichnis und Anwendung der Prozedur Univariate):
  etwa 5 Minuten zum Einlesen einer Gesamtsdatei und je Parameter weitere 5 Minuten für die Deskription, vorausgesetzt für jeden Parameter wird ein

eigener proc univariate-Aufruf angestoßen. Je Gesamtdatei sind das ca. 20 Minuten; insgesamt etwa 3 Stunden.

Mit einer Laufzeit von insgesamt ca. 20 Tagen und unter Berücksichtigung der Tatsache, dass zeitweise bis zu 14 PCs im Einsatz waren, ist der Aufwand für diese Methode sehr hoch.
Außerdem lieferte die Zensusmethode im Methodenvergleich für einige Einflussgrößen abweichende Ergebnisse in Bezug auf statistische Signifikanz und klinische Relevanz: teilweise sind die erhaltenen Gesamtbereiche sehr breit, in einem Fall traten sogar Überlappungen der Gesamtbereiche für Punktschätzer und Konfidenzgrenzen auf.

Insgesamt würde man deshalb eine der beiden anderen Methoden empfehlen, die beide mit Hilfe der SAS®-Prozeduren GENMOD [4, 5] und GLIMMIX [6] leicht und schnell durchzuführen sind. Für den Fall, dass man sich trotzdem für die Erzeugung unabhängiger Datensätze und die Anwendung gängiger Methoden entscheidet, wäre es sicher ausreichend, anstelle der Zensusmethode ein geeignetes Bootstrap-Verfahren [7] anzuwenden, das mit deutlich weniger Simulationen vergleichbare Resultate erwarten lässt.

Die Umsetzung der Zensusmethode hat gezeigt, dass SAS bereits auf Windows PCs mit sehr großen Dateien umgehen kann. Limitierender Faktor sind eher die Rechnerkapazitäten, die durch Speicher Aus- und Überlastung momentan noch zu sehr hohen Laufzeiten führen.

Literatur

[1] Wieshammer S, Dreyhaupt J: Dry Powder Inhalers: which factors determine the frequency of handling errors? Respiration. 2008; 75(1): 18-25.

[2] Dreyhaupt J, Wieshammer S, Ring C, Muche R: Vergleich verschiedener Methoden zum Anpassen eines marginalen Modells zur Schätzung der Effekte unterschiedlicher Einflussgrößen auf die ineffektive Inhalation bei Pulverinhalatoren. GMS Med. Inform. Biom. Epidemiol. (2009). (Abstract) (http://www.egms.de/en/meetings/gmds2009/09gmds107.shtml)

[3] Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130

[4] Kuss, O (2002). How to Use SAS for Logistic Regression with Correlated Data. Proceedings of the 27th Annual SAS Users Group International Conference, Paper 261-27.

[5] SUGI 26: Simplification on Learning Model by Using PROC GENMOD

[6] The GLIMMIX procedure, Juni 2006, http://support.sas.com/rnd/app/papers/glimmix.pdf

[7] Efron, B.; Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.