Publikationsfertige Kombination von Häufigkeiten und Risiko-Kennwerten aus Ergebnissen von klinisch-epidemiologischen Studien

Heribert Ramroth
Universitätsklinikum Heidelberg,
Hygiene-Institut
INF 324
69120 Heidelberg
Heribert.Ramroth@uni-heidelberg.de

Zusammenfassung

Hintergrund: Am Ende der Auswertung von klinischen und epidemiologischen Studien steht die übersichtliche Darstellung und Publikation der Ergebnisse. Die Übertragung der Ergebnisse aus den beschreibenden bzw. analytischen SAS-Prozeduren in eine publikationsfähige Tabelle erfolgt dabei oft manuell, bzw. unter manueller Nachbearbeitung des Ergebnisformats in einem Textverarbeitungsprogramm.
Ziel: Zur Erstellung der Ergebnistabelle ist es innerhalb eines SAS-Programms ohne zusätzliche manuelle Bearbeitung möglich, beschreibende und analytische Ergebnisteile durch Anwendung von SAS ODS übersichtlich zu kombinieren.
Methodik: Die Beschreibung der kategorialen Variablen des Studienkollektivs erfolgt hier mit PROC FREQ. Aus der mittels SAS ODS erzeugten Häufigkeitstabelle werden die entsprechenden absoluten Häufigkeiten mit den relativen Häufigkeiten in Klammern kombiniert, um eine sinnvolle und optisch gute Darstellung der Verteilung zu erreichen. Die zugehörige Risikoschätzung erfolgt in einem zweiten Schritt mit einer analytischen Prozedur (hier mit PROC PHREG). Die wiederum mittels SAS ODS erzeugte Ergebnistabelle lässt sich leicht optisch aufbereiten, so dass in einem letzten Schritt nun die beschreibenden und analytischen Ergebnisteile übersichtlich im SAS Programm kombiniert werden können.
Beispielhaft für drei verschiedene Arten von Variablen (binär, kategorial mit mehr als 2 Kategorien, stetig) wird ein allgemeiner Ansatz des Vorgehens vorbereitet, da die Ergebnisse von PROC FREQ und PROC PHREG in den drei Fällen auf unterschiedliche Weise kombiniert werden müssen. Anhand der Daten einer häufigkeitsgematchten Fall-Kontroll-Studie wird die Methodik demonstriert.
Schlussfolgerung: Beschreibende und analytische Prozeduren sind austauschbar, sodass sich ein allgemeiner Ansatz für andere Studienformen ableiten lässt. Bei Änderungen in den Daten oder im analytischen Modell lässt sich die Tabelle mit einem einfachen Programmdurchlauf ohne manuelle Nachbearbeitung erneut erstellen.

Schlüsselwörter: ODS; binäre, kategoriale Variable (n>2), stetige Variablen, PROC FREQ, TABULATE-Prozedur; PROC PHREG, LOGISTIC, GENMOD-Prozedur.

1 Einleitung

Am Ende der Auswertung von klinischen und epidemiologischen Studien steht die übersichtliche Darstellung und Publikation der Ergebnisse. Die Übertragung der Ergebnisse aus den beschreibenden bzw. analytischen SAS-Prozeduren in eine publikationsfähige Tabelle erfolgt dabei oft manuell bzw. unter manueller Nachbearbeitung des Ergebnisformats im verwendeten Textverarbeitungsprogramm. Ziel dieses Beitrages ist zu zeigen, dass es unter Verwendung von SAS ODS für die Darstellung der endgültigen Ergebnistabelle sinnvoll und ohne zusätzliche manuelle Bearbeitung möglich ist, beschreibende und analytische Ergebnisteile übersichtlich innerhalb eines SAS-Programms zu kombinieren.

2 Methodik

Die Methode lässt sich in drei Schritten innerhalb eines SAS-Programms strukturieren:
(i) Die Beschreibung des Studienkollektivs mittels kategorialer Variablen erfolgt zum Beispiel mit PROC FREQ (PROC TABULATE, PROC REPORT). Dabei ist sowohl die Darstellung der absoluten Häufigkeiten, als auch der relativen Häufigkeiten alleine oft unbefriedigend. Sinnvoll ist die Darstellung der Kombination von absoluten Werten mit den entsprechenden Prozentangaben in Klammern. Dazu bedarf es einer einfachen Nachbearbeitung der von PROC FREQ via SAS ODS erzeugten Ergebnisdatei.
(ii) Die zugehörige Risikoschätzung erfolgt im zweiten Schritt in einer analytischen Prozedur zum Beispiel mit PROC LOGISITIC, PROC PHREG oder PROC GENMOD. Für die endgültige publikationsfähige Darstellung der Risikokennwerte und dazugehörigen 95%-Konfidenzintervalle wird die von PROC PHREG via SAS ODS erzeugte Ergebnisdatei in einfacher Weise nachbearbeitet.
(iii) Im dritten Schritt gilt es nun, die beiden aufbereiteten beschreibenden und analytischen Ergebnisteile übersichtlich zu kombinieren.
Anhand der Daten einer häufigkeitsgematchten Fall-Kontroll-Studie wird das Vorgehen unter Verwendung von PROC FREQ und PROC PHREG demonstriert. Beispielhaft für drei verschiedene Arten von Variablen (binär, kategorial mit mehr als 2 Kategorien, stetig) wird ein allgemeiner Ansatz des Vorgehens vorbereitet, da die Ergebnisse von PROC FREQ und PROC PHREG bei diesen drei Arten von Variablen auf unterschiedliche Weise kombiniert werden müssen.
Beschreibende und analytische Prozeduren sind austauschbar, sodass sich ein allgemeiner Ansatz für andere Studienformen ableiten lässt.

2.1 Voraussetzungen

Unterschieden wird in der Praxis im Allgemeinen zwischen kategorialen und stetigen Variablen, wobei der Sonderfall der binären Variablen aus programm-technischen Gründen hier gesondert betrachtet wird. Der folgende Text unterscheidet somit binäre Variablen, kategoriale nicht binäre Variablen (mit mehr als 2 Kategorien) und stetige Variablen. Die stetigen Variablen spielen hier nur eine marginale Rolle, da sie in das analytische Modell eingehen, jedoch nicht bei der Häufigkeitsauszählung auftauchen.

Das gleiche gilt für kategoriale Variablen, die zur Trendanalyse ins analytische Modell eingehen, ohne Klassifizierung mittels *class*-statement.
Zur Veranschaulichung des Vorganges werden hier je eine binäre, eine kategoriale und eine stetige Variable verwendet.

Binäre Variable
Es soll hier vorausgesetzt werden, dass die binären Variablen nur mit 0/1 kodiert werden. Sollte dies nicht gegeben sein, muss vorab eine entsprechende Umkodierung durchgeführt werden.

Kategoriale Variable
Eine kategoriale Variable im hier verwendeten Sinne hat also mindestens 3 Kategorien. Im folgenden soll hier ohne Beschränkung der Allgemeinheit vorausgesetzt werden, dass diese mit den Einträgen 1, 2, 3, ... usw. kodiert werden. Insbesondere bedeutet dies, dass die kleinste Kategorie nicht mit „0“ kodiert wird. Im Einklang mit dem hier aufgezeigten Beispiel sei angenommen, dass die hier verwendete kategoriale Variable 4 Kategorien hat. Für manche analytischen Anwendungen wie auch PROC PHREG ist es notwendig, diese kategoriale Variable mit n Kategorien in n Dummy Variablen umzukodieren, wobei im analytischen Modell unter Verzicht auf die Referenzkategorie nur n-1 Dummy-Variablen verwendet werden. Bei der ausschließlichen Verwendung von Dummy-Variablen fällt somit die kategoriale Variable wieder auf den Fall der binären Variablen zurück. Die kategoriale Variable selbst lässt sich im analytischen Modell zwar zur Trendanalyse verwenden, wobei ihr in diesem Fall jedoch keine Häufigkeiten gegenübergestellt werden.

2.2 Beispiel

Das Anwendungsbeispiel verwendet Daten einer deutschen Fall-Kontrollstudie zum Kehlkopfkarzinom [1]. Die verwendeten Variablen dienen dabei der Beschreibung der Methodik des Themas und geben nicht die beste Modellanpassung wieder.

Variable	Name	Label	Kodierung	Format
Binär	JsR	Jahre seit Rauchende	0, 1	0: 0-1 Jahre, 1: >=2 Jahre
Kategorial	pck4kat	Packungsjahre des Rauchens (PJ)	1, 2, 3, 4	1=0 2=0 < PJ < 20 3=20 <= PJ < 40 4=40 <= PJ < ..
Stetig	EthTag	Ethanol pro Tag	-	g Ethanol / 25

Für das Analysemodell werden aus der kategorialen Variablen pck4kat die 3 Dummy-Variablen pck4kat2, pck4kat3 und pck4kat4 konstruiert

Variable	Name	Label	Kodierung	Format
Binär	pck4kat2	0 < PJ < 20	0, 1	1=(0 < PJ < 20) / 0=sonst
Binär	pck4kat3	20 <= PJ < 40	0, 1	1=(20 <= PJ < 40) / 0=sonst
Binär	pck4kat4	20 <= PJ < 40	0, 1	1=(40 <= PJ < ..) / 0=sonst

Zur Auswertung dieser häufigkeitsgematchten Fall-Kontroll-Studie mittels bedingter logistischer Regression wird die Prozedur PROC PHREG verwendet. Details (insbesondere auch für die Notwendigkeit der Konstruktion der künstlichen Zensierungsvariablen *Time* sind z.B. nachzulesen bei Hosmer & Lemeshow 1989 [2] bzw. in der SAS-Hilfe (SAS-Example 54.3: Conditional Logistic Regression for m:n Matching [3]).

Im Beispiel werden folgende weitere Variablen benötigt:

Fallkontrollstatus, *caco*, binär (0=Kontrollen, 1=Fälle)
Stratifizierungsvariable, *SexAgeGrp*, kategorial (Altersgruppe 1 bis n)
Dummy-Zensierungsvariable, *Time*, binär (2=Kontrollen, 1=Fälle).

3 Deskriptive Statistik

Die bekannteste Prozedur zur Angabe von absoluten und relativen Häufigkeiten ist sicherlich PROC FREQ. Dem Vorteil der einfachen Syntax dieser Prozedur steht jedoch der Nachteil gegenüber, das die Ergebnisdarstellung für die direkte Übernahme in einer Publikation wenig ansprechend ist (Tabelle 1).

```
proc freq data=smoke;
tables JsR * caco /nopercent norow;
run;
```

Tabelle 1: Ausgabe von PROC FREQ der Variablen JsR im OUTPUT–Fenster

Table of JsR by caco

JsR (Jahre seit RauchEnde)		*Caco*		
Frequency *Col Pct*		*0*	*1*	*Total*
	0	383 49.80	176 68.48	559
	1	386 50.20	81 31.52	467
Total		769	257	1026

Die Darstellung der Häufigkeitsverteilung mit PROC TABULATE kommt dem gewünschten Endergebnis zwar am nächsten (Tabelle 2), es ist jedoch nicht möglich auf genau diese Form der Ergebnisausgabe zuzugreifen. Dies wäre aber für die Kombination mit den Risikokennwerten wichtig.

```
proc tabulate data=smoke format=5.1 ;
class JsR pck4kat caco;
     table JsR pck4kat, caco*(n*f=5. colpctn='(%)');
run;
```

Tabelle 2: Ausgabe von PROC TABULATE der absoluten und relativen Häufigkeiten der binären und kategorialen Variablen im OUTPUT-Fenster

	Caco			
	Kontrollen		Fälle	
	N	(%)	N	(%)
Jahre seit Rauchende				
0-1	383	49.8	176	68.5
2+	386	50.2	81	31.5
4 Packungsjahr-Gruppen				
0	203	26.4	9	3.5
0< PJ < 20	297	38.6	33	12.8
20<= PJ < 40	147	19.1	88	34.2
40<= PJ < ...	122	15.9	127	49.4

Unabhängig davon, ob die Häufigkeiten mittels PROC FREQ oder PROC TABULATE ausgezählt werden, lassen sich bei beiden Prozeduren die absoluten und die relativen Häufigkeiten mittels SAS ODS in einer Ausgabe-Tabelle abspeichern. Dazu wird der entsprechende Ergebnisteil der Procedur mittels SAS statement *ods trace on* identifiziert und mittels *ods output* an eine Ausgabedatei weitergegeben. Da sich diese Ausgabedatei für PROC FREQ bzw. PROC TABULATE nur unwesentlich unterscheidet, wird das Vorgehen aufgrund der einfachen Syntax anhand der Prozedur PROC FREQ dargestellt. Überflüssige Prozentangaben sind im folgenden Prozedurschritt weggelassen, da in einer Fall-Kontroll-Studie die Prozentangaben der Exposition in der jeweiligen Gruppe von Fällen bzw. Kontrollen von Interesse sind.

```
ods trace on / listing;
ods output CrossTabFreqs=ctf;
proc freq data=smoke;
tables (JsR pck4kat) * caco /nopercent norow;
run;
ods trace off;
```

ods trace on identifiziert den Ergebnisteil *CrossTabFreqs*, der bei einer Kreuztabelle von 2 oder mehr Variablen (hier z.B. JsR * caco) die Häufikeiten enthält. Die option *listing* gibt den Namen des Ergebnisteils im Output-Fenster aus, statt im Log-Fenster. Mittels *ods output* werden Häufigkeiten und Informationen aus der PROC FREQ an eine selbst benannte Ausgabedatei weitergegeben, die hier *ctf* genannt wird.
Die Informationen der Datei *ctf* sind in Tabelle 3 wiedergegeben. Automatisch werden in der Dabei ctf zusätzlich zu den Häufigkeiten und Prozentangaben die Variablen *Table*, *_TYPE_* und *missing* erzeugt. Aus Platzgründen ist in Tabelle 3 nur der Ergebnisteil für die Variable *JsR* dargestellt. Die Informationen bzgl. der Variablen *pck4kat* sind jedoch auch in weiteren Zeilen dieser Tabelle enthalten.
Die Variable Table enthält die Information der Variablen, von welcher die Häufigkeiten in der Zeile folgen. Die Textvariable _TYPE_ zeigt an, ob es sich hier um innere Werte der Kreuztabelle handelt (_TYPE_ ="11"), bzw. die Randverteilungen bzgl. der Variablen JsR (_TYPE_ ="10") bzw. caco (_TYPE_ ="01"). Die Summe der eingegangenen Beobachtungen ist mit _TYPE_ ="00" gekennzeichnet. Dies lässt sich leicht anhand der beiden Tabellen 1 und 3 nachvollziehen. Die Variable JsR enthält die Werte der Kodierung der Originalvariablen. Für jede zusätzliche Variable im *tables statement* der Variablenliste von PROC FREQ ist eine weitere Spalte in der Tabelle ctf mit dem entsprechenden Wertebereich vorhanden. Bei Auftreten von fehlenden Werten ist diese Information in der Spalte *Missing* abgelegt.

```
proc print data=ctf;
run;
```

Tabelle 3: Von PROC FREQ erzeugte Ergebnistabelle ctf (Variable JsR)

Obs	Table	caco	_TYPE_	Frequency	ColPercent	JsR	pck4kat	Missing
1	Table JsR * caco	0	11	383	49.8049	0	.	.
2	Table JsR * caco	1	11	176	68.4825	0	.	.
3	Table JsR * caco	.	10	559	.	0	.	.
4	Table JsR * caco	0	11	386	50.1951	1	.	.
5	Table JsR * caco	1	11	81	31.5175	1	.	.
6	Table JsR * caco	.	10	467	.	1	.	.
7	Table JsR * caco	0	01	769	.	.	.	.
8	Table JsR * caco	1	01	257	.	.	.	.
9	Table JsR * caco	.	00	1026	.	.	.	0
	.....							

Mit einem einfachen Datenschritt lassen sich nun absolute und relative Häufigkeiten optisch ansprechend in einer neuen Textvariablen *NPct* kombinieren.

```
data ctf1;
set ctf;
NPct=Frequency||" ("||PUT(ColPercent,4.1)||")";

... datastep wird fortgesetzt ...
```

Verwendung von PUT zur Formatzuweisung ergibt eine Dezimalstelle für jede Prozentangabe, auch im Falle von ganzen Zahlen. Häufigkeiten und formatierte Prozentangaben werden nun noch mittels Klammern und führendem Leerzeichen kombiniert. Die folgende Tabelle 4 zeigt die Ausgabe von PROC PRINT nur der inneren Werten der Kreuztabelle JSR*caco (d.h. _TYPE_="11") und der neu konstruierten Variablen *NPct*.

Tabelle 4: Erzeugung der Textvariablen NPct aus Frequency und ColPercent.

Obs	Variable	Table	caco	Frequency	ColPercent	JsR	pck4kat	NPct
1	JsR	Table JsR * caco	0	383	49.8049	0	.	383 (49.8)
2	JsR	Table JsR * caco	1	176	68.4825	0	.	176 (68.5)
4	JsR	Table JsR * caco	0	386	50.1951	1	.	386 (50.2)
5	JsR	Table JsR * caco	1	81	31.5175	1	.	81 (31.5)

Auf eine Darstellung der Werte bzgl. der Variablen pck4kat entsprechend den Tabellen 2 bis 4 wird hier aus Übersichtsgründen verzichtet.

Das Ziel einer Darstellung der Häufigkeiten von Fällen und Kontrollen in 2 Spalten erfordert die Anwendung von PROC TRANSPOSE auf die inneren Werte von Tabelle ctf1. Zum Transponieren der entsprechenden Werte wird aus den Namen (*Variable*) und den Werten (*Value*) der ursprünglichen Variablen eine eindeutige Variable TranspVar konstruiert. Da der Wertebereich der ursprünglichen Variablen JsR und pck4Kat in 2 verschiedenen Spalten gespeichert ist, lässt sich TranspVar leicht unter Verwendung eines Arrays erzeugen. Die Ergebnisdatei ctf1 ist in Tabelle 5 wiedergegeben.

```
... Fortsetzung des datasteps...
Variable=scan(Table, 2);
array ArrVars JsR pck4kat;
do over ArrVars;
   if ArrVars ne . then do;
   value=put(ArrVars,1.);
   TranspVar=cats(variable,value);
   end;    *Ende der Do-Schleife;
end;    *Ende des do-over-Array-Schrittes;
run;    *Ende des Datenschrittes;
```

Tabelle 5: TranspVar, erzeugt aus den originalen Variablennamen und deren Werten

Obs	Table	caco	NPCT	JsR	pck4kat	Variable	value	TranspVar
1	Table JsR * caco	0	383 (49.8)	0	.	JsR	0	JsR0
2	Table JsR * caco	1	176 (68.5)	0	.	JsR	0	JsR0
4	Table JsR * caco	0	386 (50.2)	1	.	JsR	1	JsR1
5	Table JsR * caco	1	81 (31.5)	1	.	JsR	1	JsR1
10	Table pck4kat * caco	0	203 (26.4)	.	1	pck4kat	1	pck4kat1
11	Table pck4kat * caco	1	9 (3.5)	.	1	pck4kat	1	pck4kat1
13	Table pck4kat * caco	0	297 (38.6)	.	2	pck4kat	2	pck4kat2
14	Table pck4kat * caco	1	33 (12.8)	.	2	pck4kat	2	pck4kat2
16	Table pck4kat * caco	0	147 (19.1)	.	3	pck4kat	3	pck4kat3
17	Table pck4kat * caco	1	88 (34.2)	.	3	pck4kat	3	pck4kat3
19	Table pck4kat * caco	0	122 (15.9)	.	4	pck4kat	4	pck4kat4
20	Table pck4kat * caco	1	127 (49.4)	.	4	pck4kat	4	pck4kat4

Durch die Konstruktion dieser eindeutigen Variablen TranspVar ist nun PROC TRANSPOSE anwendbar. Mittels p*refix=caco* und *id caco* werden die neuen beiden Spalten *caco0* und *caco1* benannt. Transponiert werden dabei nur innere Werte der Kreuztabelle ctf1 (_TYPE_="11"). Die Ausgabe erfolgt in Datei ctf2 (siehe Tabelle 6).

```
proc transpose data=ctf1 out=ctf2 prefix=caco;
var NPct;
id caco;
by TranspVar;
where _TYPE_="11";
proc print data=ctf2;
run;
```

Tabelle 6: Die Ergebnisdatei ctf2 ausgegeben von PROC TRANSPOSE.

Obs	TranspVar	caco0	caco1
1	JsR0	383 (49.8)	176 (68.5)
2	JsR1	386 (50.2)	81 (31.5)
3	pck4kat1	203 (26.4)	9 (3.5)
4	pck4kat2	297 (38.6)	33 (12.8)
5	pck4kat3	147 (19.1)	88 (34.2)
6	pck4kat4	122 (15.9)	127 (49.4)

Somit ist Teil 1 der gestellten Aufgabe, die Kombination von absoluten und relativen Häufigkeiten und die Ausgabe in einer weiter verarbeitbaren Datei, erfüllt.

4 Analytische Statistik

Die Verwendung von PROC PHREG zur Auswertung einer häufigkeitsgematchten Fall-Kontrollstudie kann am Beispiel Hosmer & Lemeshow nachgeschlagen werden [2].
Als Modellvariablen werden hier die binäre Variable JsR, die kategoriale Variable pck4kat und die stetige Variable EthTag ins Modell aufgenommen. Aus der kategorialen Variablen pck4kat wurden die 3 Dummy-Variablen pck4kat2-pck4kat4 mit Referenzkategorie pck4kat1 (Nichtraucher, äquivalent 0 Packungsjahre) erzeugt.
Die künstliche Dummy-Zensierungsvariable Time wurde entsprechend Kapitel 2 (vgl. Referenz [3]) konstruiert. Die Variable SexAgeGr enthält die Gruppierung in 5-Jahres-Alters-und-Geschlechtsgruppen entsprechend den Matching-Kriterien der Fall-Kontroll-Studie. Analog Kapitel 3 wird der Ergebnisteil von PROC PHREG, welcher die Risikokennwerte enthält (*ParameterEstimates*) im Bedarfsfall mittels *ods trace on* identifiziert und mittels *ods output* an eine selbst benannte Ausgabedatei mit Namen *ParmEst* weitergegeben.

```
ods output ParameterEstimates=ParmEst;
proc phreg data = Smoke;
model Time*caco(0)= JsR pck4kat2 pck4kat3 pck4kat4 EthTag /
ties=discrete;
strata SexAgeGr;
```

```
proc print data=ParmEst;
run;
```

Die Ausgabe von PROC PHREG ist in Tabelle 7 wiedergegeben. Dabei sind die SAS Label der Variablen (JsR, pck4kat2 - pck4kat4, EthTag) hier aus Platzgründen nicht angezeigt; sie sind jedoch in der Datei *ParmEst* als Variable *Label* gespeichert. Auf eine Interpretation der Ergebnisse wird hier bewusst verzichtet.

Tabelle 7: Ausgabe von PROC PHREG im OUTPUT-Fenster

Variable	DF	Estimate	StdError	ChiSq	Prob ChiSq	Hazard Ratio	HR LowerCL	HR UpperCL
JsR	1	-1.04362	0.19108	29.8314	<.0001	0.352	0.242	0.512
pck4kat2	1	1.79274	0.41733	18.4535	<.0001	6.006	2.651	13.608
pck4kat3	1	3.13383	0.39288	63.6265	<.0001	22.962	10.631	49.593
pck4kat4	1	3.44666	0.38944	78.3283	<.0001	31.395	14.634	67.353
EthTag	1	0.00560	0.00133	17.6387	<.0001	1.150	1.077	1.228

Die Ausgabe der Odds Ratios und 95%-Konfidenzintervalle der Prozedur PHREG wird mit einfachen Mitteln durch Aufbereitung der Datei *ParmEst* optisch verbessert und in der Ergebnisdatei *ParmEst1* ausgegeben (Tabelle 8).

```
data ParmEst1;
set ParmEst;
OR=PUT(HazardRatio, 5.1);
CI=CATS("(",PUT(HRLowerCL,5.1))||", "||CATS(PUT(HRUpperCL,5.1),")");

proc print data=ParmEst1;
run;
```

Mit der PUT-Funktion lässt sich ein Format mit einer Dezimalstelle erreichen (Der Unterschied gegenüber ROUND besteht darin, dass auch ganzzahlige Ergebnisse wie 6.0 auf 1 Dezimalstelle dargestellt werden, und nicht als „6“.). Die Funktion CATS fügt die Klammer und die untere Konfidenzgrenze ohne Leerstellen zusammen (analog die obere Konfidenzgrenze und die schließende Klammer). Dazwischen wir nun noch ein Komma mit Leerzeichen als optisches Trennzeichen eingefügt.

Tabelle 8: Aufbereitete Odds Ratios (OR) mit den entsprechenden 95%-Konfidenzintervallen (CI) in der Datei *ParmEst1*

Obs	Variable	Hazard Ratio	HR Lower CL	HR Upper CL	Label	OR	CI
1	JsR	0.352	0.242	0.512	Jahre seit RauchEnde	0.35	(0.24, 0.51)
2	pck4kat2	6.006	2.651	13.608	0 < PJ < 20	6.0	(2.7, 13.6)
3	pck4kat3	22.962	10.631	49.593	20 <= PJ < 40	23.0	(10.6, 49.6)
4	pck4kat4	31.395	14.634	67.353	40 <= PJ	31.4	(14.6, 67.4)
5	EthTag	1.150	1.077	1.228	g Ethanol pro Tag/25	1.2	(1.1, 1.2)

Der Unterschied zwischen dem obigen einfachen Programmcode zur Aufbereitung der Odds Ratios (OR) und den 95%-Konfidenzintervallen in der Datei *ParmEst1* und dem in Tabelle 8 dargestellten Ergebnis besteht darin, das für die exakte Ausgabe von Tabelle 8 noch einige wenige extra Programmzeilen nötig sind. Diese tragen jedoch nur der Konvention Rechnung, Dezimalzahlen größer als 1 mit nur 1 Dezimalstelle darzustellen, Dezimalzahlen kleiner als 1 dagegen mit 2 Dezimalstellen. Die Zeile zur optischen Verbesserung der Odds Ratios (OR) wird in diesem Falle wie folgt ersetzt:

```
if HazardRatio<1 then OR=PUT(HazardRatio,5.2);
   else OR=PUT(HazardRatio,5.1);
```

Die Zeile zur optischen Verbesserung der Konfidenzintevalle (ci) wird wie folgt ersetzt:

```
select;
when(LCL<1 and UCL<1)
          ci= CATS("(",PUT(LCL,5.2))||", "|| CATS(PUT(UCL,5.2),")");
when(LCL<1 and UCL>=1)
          ci= CATS("(",PUT(LCL,5.2))||", "|| CATS(PUT(UCL,5.1),")");
when(LCL>=1 and UCL>=1)
          ci= CATS("(",PUT(LCL,5.1))||", "|| CATS(PUT(UCL,5.1),")");
otherwise;
end;
```

Somit ist Teil 2 der gestellten Aufgabe, die Aufbereitung und Ausgabe der Risikokennwerte in einer weiter verarbeitbaren Datei, erfüllt.

5 Kombination der Ergebnisse deskriptiver und analytischer Statisik

In wenigen Schritten lassen sich nun die beiden Ergebnisteile aus Kapitel 3 und 4 kombinieren. Ausgangspunkte sind die oben konstruierten Tabellen 6 mit den Häufigkeitsverteilungen und Tabelle 8 mit den Risikokennwerten.

Die Variablen *TranspVar* (Tabelle 6) und *Variable* (Tabelle 8) enthalten die Werte, die für eine Kombination (*Merge*) der beiden Tabellen wichtig sind. Betrachtet man die in Tabelle 6 konstruierte Variable TranspVar, so ist nur für die binäre Variable *JsR* eine unterschiedliche Notation der Spalteneinträge festzustellen. Die Häufigkeits-Information, die in Tabelle 6 für den Zeileneintrag „JsR1“ gespeichert ist, muss mit den Risiko-Kennwerten für den Zeileneintrag „JsR“ aus Tabelle 8 *gemergt* werden. Hier sind sicherlich verschiedene Varianten möglich. Die hier vorgeführte Variante bedient sich der Tatsache, das laut Voraussetzung (Kapitel 2) nur bei binären Variablen der Werteeintrag „0“ zulässig ist, nicht jedoch bei kategorialen Variablen mit mehr als 2 Kategorien.
Mittels der Funktion *vlag* wird nun eine Variable *MergeVar* konstruiert, die bis auf den Sonderfall bei binären Variablen mit der Transpositionsvariablen TranspVar identisch ist. Um zu gewährleisten, dass es aufgrund von nicht besetzten Zellhäufigkeiten zu Fehlern beim Transponieren kommt, setzt die Konstruktion von MergeVar schon bei der in Kapitel 3 erzeugten Tabelle ctf1 mit der Randverteilung (_TYPE_=“10“) an. Als Hilfe bei evtl. nötigen Sortierungen wird hier zusätzlich eine Sortiervariable MyOrder konstruiert, die später eine Darstellung von genau der Reihenfolge der Beobachtungen zulässt, wie sie ursprünglich im *tables* statement der PROC FREQ eingegeben wurden.

Tabelle 9: Hilfstabelle zur Kombination der Ergebnisse der beschreibenden und analytsichen Ergebnisse.

Obs	Variable	value	TranspVar	MyOrder	vlag	MergeVar
1	JsR	0	JsR0	1		JsR0
2	JsR	1	JsR1	2	0	JsR
3	pck4kat	1	pck4kat1	3	1	pck4kat1
4	pck4kat	2	pck4kat2	4	1	pck4kat2
5	pck4kat	3	pck4kat3	5	2	pck4kat3
6	pck4kat	4	pck4kat4	6	3	pck4kat4
7	EthTag	1	EthTag	7	4	EthTag

```
Data MergeTable;
     set ctf1 (keep= Variable Value TranspVar);
     where _TYPE_="10" ;
     vlag=lag(Value);
     if vlag=0 then MergeVar=Variable;
                else MergeVar=TranspVar;
     MyOrder=_N_;
run;
proc print data=MergeTable;
run;
```

Da die beiden Merge-Variablen in Tabelle 6 (*TranspVar*) und 8 (*Variable*) unterschiedliche Namen haben, bietet sich hier PROC SQL zur Kombination der Informationen aus beiden Tabellen an [4]. Tabelle 8 enthält zusätzlich die stetige Variable EthTag aus dem analytischen Modell.

```
proc sql;
   create table ResSmoke as
   select variable, value, caco0, caco1, or, ci format $12., label
   from ctf2 as f full join
   (select s.variable as variable, MergeVar, TranspVar, MyOrder,
    value, or, ci, label
   from ParmEst1 as o full join MergeTable as s
   on s.MergeVar=o.Variable) as s
   on s.TranspVar=f.TranspVar
   order by MyOrder;
quit;
```

Tabelle 10: Ergebnistabelle mit absoluten und relativen Häufigkeiten, sowie Risikokennwerten

Obs	Variable	value	caco0	caco1	OR	ci	Label
1	JsR	0	383 (49.8)	176 (68.5)			
2	JsR	1	386 (50.2)	81 (31.5)	0.35	(0.24, 0.51)	Jahre seit RauchEnde
3	pck4kat	1	203 (26.4)	9 (3.5)			
4	pck4kat	2	297 (38.6)	33 (12.8)	6.0	(2.7, 13.6)	0 < PJ < 20
5	pck4kat	3	147 (19.1)	88 (34.2)	23.0	(10.6, 49.6)	20 <= PJ < 40
6	pck4kat	4	122 (15.9)	127 (49.4)	31.4	(14.6, 67.4)	40 <= PJ
7	EthTag				1.1	(1.1, 1.2)	g Ethanol pro Tag/25

Zur optischen Verfeinerung lassen sich mit Labels die Variablen caco0 (=Kontrollen), caco1 (=Fälle) und ci (=95%CI) natürlich noch beschriften. Per Programm kann in der Referenzkategorie zusätzlich ein OR von 1 und ci="-„ eingetragen werden.

6 Schlussfolgerung

Das oben gezeigte Verfahren orientiert sich an den Ergebnisdateien der deskriptiven und der analytischen Prozeduren. Zwischen diesen beiden Ergebnisdateien wird dann eine Brücke gebaut in Form einer verbindenden Tabelle (hier: MergeTable). Bei Verwendung anderer deskriptiver und/oder analytischer Prozeduren ist es daher leicht möglich, eine entsprechend angepasste Verbindungstabelle zu konstruieren. Das Verfahren ist somit leicht auf andere Auswertungssituationen in klinischen und epidemiologischen Studien anwendbar. Da keine der Werte innerhalb der Tabelle im Textverarbeitungsprogramm nachbearbeitet wurden, lässt sich diese Tabelle mit einem einfachen Programmdurchlauf wiederholen, sollten sich Änderungen in den Daten oder im analytischen Modell ergeben.

Literatur

[1] Ramroth H, Dietz A, Becher H. Interaction effects and population-attributable risks for smoking and alcohol on laryngeal cancer and its subsites. Methods Inf Med 2004; 43 (5), 499-504

[2] Hosmer DWJ & Lemeshow S (1989) *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

[3] SAS-Example 54.3: Conditional Logistic Regression for m:n Matching.

[4] SAS Institute Inc., *Getting started with the SQL-Procedure*, *Version 6, First Edition*, Cary, NC: SAS Institute Inc., 1994, 78pp.