Familiäres Krebsrisiko: Berechnung relativer Risiken in SAS

Christoph Frank
Deutsches Krebsforschungszentrum
Abteilung für Molekulargenetische Epidemiologie
Im Neuenheimer Feld 580
69120 Heidelberg
Christoph.frank@dkfz.de

Zusammenfassung

Angehörige von Krebspatienten sind selbst einem erhöhten Krebsrisiko ausgesetzt. Heute bekannte genetische Veranlagungen erklären jedoch nur einen kleinen Teil des familiären Krebsaufkommens. Um die Krankheitsursachen zu erforschen und um verlässliche Risikoschätzungen für Angehörige von Patienten bereitzustellen, wurden relative Risiken für verschiedene Verwandtschaftsverhältnisse mit Hilfe eines Poisson Regressionsmodells basierend auf Daten der schwedischen Familien-Krebs-Datenbank geschätzt.

Schlüsselwörter: Epidemiologie, familiärer Krebs, Relatives Risiko, Poisson Regression, PROC GENMOD

1 Einleitung

Die Volkskrankheit Krebs war, entsprechend des kürzlich erschienen World Cancer Report 2014 der International Agency for Research on Cancer (IARC), die weltweit häufigste Todesursache im Jahre 2012 gefolgt von koronarer Herzerkrankung und Schlaganfall. Schätzungsweise 14.1 Millionen Neuerkrankungen und 8.2 Millionen krebsbedingte Todesfälle wurden 2012 registriert.

Es ist wohlbekannt, dass enge Verwandte von Krebspatienten einem erhöhten Risiko ausgesetzt sind, an derselben Krebsart zu erkranken. Treten mehrere gleichartige Krebserkrankungen innerhalb einer Familie auf, spricht man von familiärem Krebs. Heute bekannte genetische Veranlagungen, auf die Angehörige von Patienten im Rahmen von genetischer Beratung getestet werden können, sind zu einem kleinen Teil verantwortlich für erhöhtes familiäres Krebsaufkommen. Für die überwiegende Mehrheit von familiären Krebsfällen sind die Ursachen jedoch weiterhin ungeklärt. Für die Beratung von Krebspatienten und deren Angehörigen ist die Bereitstellung von verlässlichen Risikomaßen für verschiedene familiäre Vorgeschichten daher von großer Bedeutung, um Ängste zu nehmen und Präventionsstrategien zu entwickeln.

Basierend auf Daten der weltweit größten Populationsdatenbank dieser Art, der schwedischen Familien-Krebs-Datenbank, wurden mit Hilfe des vorliegenden SAS-Programms relative Risiken im Rahmen eines Poisson Regressionsmodells geschätzt und miteinander verglichen.

2 Poisson Regression – Theoretische Grundlage

Es ist ein altbewährter epidemiologischer Ansatz, Risiken für Krankheiten unter Berücksichtigung bestimmter Risikofaktoren zu studieren, um die Ätiologie bzw. die Ursachen von Krankheiten zu erforschen. Insbesondere das relative Risiko (RR) stellt ein oft verwendetes Maß dar, um den Effekt eines Risikofaktors zu bewerten.
Wenn bei der Berechnung von RRs weitere potenzielle Risikofaktoren (Confounders) miteinbezogen werden sollen, um verzerrte Schätzungen zu vermeiden, stellen generalisierte lineare Modelle (GLMs) ein probates Mittel dar. Dabei wird die Beziehung zwischen einer abhängigen Variable Y (Response) und unabhängigen Variablen X (Kovariaten) unter der Annahme modelliert, dass Y eine Verteilung aus der Klasse der Exponentialfamilie aufweist, deren Verteilungsparameter durch eine Link-Funktion vom linearen Prädiktor Xβ bestimmt werden. Die Regressionskoeffizienten β sind Maximum-Likelihood-Schätzer. Das populärste GLM zur Modellierung von Zähldaten ist Poisson Regression.
In einem Poisson Regressionsmodell wird angenommen, dass die diskrete Response-Variable Y Poisson-verteilt ist und folgende Wahrscheinlichkeitsfunktion aufweist:

$$f(y_i; \mu_i) := P(Y = y_i) = \frac{e^{-\mu_i}\, {\mu_i}^{y_i}}{y_i!}\,, \qquad y_i = 0,1,2,\ldots; \mu_i > 0\,.$$

Dabei entspricht y_i der beobachteten Anzahl bzw. den beobachteten Werten von Y. Der Index i verweist auf das Modellieren von mehreren Beobachtungen y_i $(i = 1, \ldots, N)$ mit Hilfe von K Kovariaten ${x_i}' = (1, x_{i1}, \ldots, x_{iK})$. Ein charakteristisches Merkmal der Poisson Verteilung ist, dass der Erwartungswert μ_i gleich der Varianz der Verteilung ist (Equidispersion). Entsprechend der Regressionsgleichung

$$\ln \mu_i = {x_i}'\beta,$$

ist μ_i durch den natürlichen Logarithmus als Link-Funktion mit dem linearen Prädiktor ${x_i}'\beta$ assoziiert. Die Regressionskoeffizienten β werden derart gewählt, dass sie die Log-Likelihood-Funktion

$$l(\beta|X, Y) := \sum_{i=1}^{N} \left[y_i({x_i}'\beta) - e^{{x_i}'\beta} - \ln y_i!\right]$$

maximieren. Die GENMOD Prozedur, mit der GLMs in SAS implementiert werden können, nutzt einen Newton-Raphson-Algorithmus, um $l(\beta|X, Y)$ zu maximieren [1].

3 Poisson Regression – Bestimmung relativer Risiken

Für die Berechnung von relativen familiären Krebsrisiken wurde der obige Ansatz folgendermaßen modifiziert. Gegeben seien y_i als die Anzahl der Neuerkrankungen, die während n_i Personenjahren (entspricht der Beobachtungszeit) diagnostiziert wurden, wobei $i = 1, \ldots, N$ Beobachtungen für jeweils voneinander verschiedene Kombinationen x_i der Kovariaten vorliegen. Neben der für diese Studie primär interessanten Informa-

tion über die familiäre Vorgeschichte, wurden zur Verfügung stehende Informationen über Altersgruppe, Geschlecht, Dekade, sozioökonomischen Index (Beruf, Einkommen, Bildung) und Wohnregion als Kovariaten betrachtet. Diese potenziellen Confounder wurden zur Adjustierung der Ergebnisse miteinbezogen.
Die Anwendung eines Poisson Regressionsmodels in diesem Kontext impliziert die Modellierung von Inzidenzraten y_i/n_i. Unter der Annahme, dass y_i Poisson verteilt ist mit Erwartungswert μ_i, folgt

$$E\left(\frac{y_i}{n_i}\right) = \frac{\mu_i}{n_i}.$$

Die Regressionsgleichung lautet dann

$$\ln\left(\frac{\mu_i}{n_i}\right) = x_i'\beta \qquad \text{bzw.} \qquad \ln\mu_i = x_i'\beta + \ln n_i\,,$$

wobei $\ln n_i$ als Offset bezeichnet wird. Für die geschätzten Inzidenzraten gilt:

$$\frac{\mu_i}{n_i} = e^{x_i'\beta}$$

Jede Kombination von Kovariaten $x_i' = (1, x_{i1}, \dots, x_{iK})$ enthält unter anderem die Information über familiäre Vorgeschichte $x_{i*} \in (x_{i1}, \dots, x_{iK})$. Wenn lediglich zwischen positiver familiärer Vorgeschichte (Krebspatienten in der Familie) und negativer familiärer Vorgeschichte (keine Krebspatienten in der Familie) unterschieden wird, dann ist x_{i*} eine binäre Variable, die derart definiert wird, dass $x_{i*} = 0$ ist bei negativer familiärer Vorgeschichte und $x_{i*} = 1$ ist bei positiver familiärer Vorgeschichte. Der zugehörige Regressionskoeffizient sei mit β^* bezeichnet. Es werden nun mit Ausnahme von x_{i*} die Werte aller Kovariaten aus x_i als konstant angenommen, wobei $(\mu_i/n_i)^+$ der Inzidenzrate für $x_{i*} = 1$ entspricht und $(\mu_i/n_i)^-$ die entsprechende Rate für $x_{i*} = 0$ darstellt. Der Quotient aus beiden Inzidenzraten ist unabhängig von i und ist ein Maß für das relative Risiko bezüglich des Risikofaktors familiäre Vorgeschichte:

$$RR = \frac{(\mu_i/n_i)^+}{(\mu_i/n_i)^-} = e^{\beta^*}.$$

Wenn mehrere verschiedene familiäre Vorbelastungen (z.B. betroffenes Elternteil, Geschwister, mehrere betroffene Verwandte) simultan analysiert werden sollen, ist x_{i*} eine kategorische Variable mit mehr als zwei Werten. RRs können dann analog wie im obigen binären Fall berechnet werden, wenn x_{i*} mit Hilfe von Dummy Variablen durch Referenz-Kodierung dargestellt wird (siehe Tabelle 1).

Innerhalb der GENMOD Prozedur lassen sich RRs mit Hilfe des ESTIMATE Befehls berechnen. Wald-Schätzer erlauben zudem Hypothesentests, um die Signifikanz von familiärer Vorgeschichte als Risikofaktor zu bewerten und ermöglichen die Berechnung von Konfidenzintervallen für RRs. Für ein gegebenes Signifikanzniveau α (z.B. 5%), ist familiäre Vorgeschichte genau dann ein statistisch signifikanter Risikofaktor, wenn 1 außerhalb des $(1-\alpha)$-Konfidenzintervalls des RRs liegt. Darüber hinaus lassen sich mittels des CONTRAST Befehls weitere Hypothesen testen. Um Rückschlüsse über

mögliche genetische Vererbungsmuster zu gewinnen, kann beispielsweise getestet werden, ob sich RRs für verschiedene Verwandtschaftsverhältnisse signifikant voneinander unterscheiden. Außerdem können Interaktionen zwischen familiärer Vorgeschichte und anderen Kovariaten untersucht werden.

Tabelle 2: Referenz-Kodierung für die Analyse verschiedener Verwandtschaftsverhältnisse

Familiäre Vorgeschichte	x_{i*}	Dummy-Variablen	
		$x_{i*}{}^{1}$	$x_{i*}{}^{2}$
Keine betroffenen Verwandten	0	0	0
Betroffenes Elternteil	1	1	0
Betroffene(r) Bruder/Schwester	2	0	1

Die immanente Annahme von Equidispersion bei Poisson Regressionsmodellen wird von empirischen Daten nur selten erfüllt. Dies lässt sich anhand der Deviance oder der Pearson Chi-Quadrat Goodness of Fit Statistik und deren zugehörigen Freiheitsgraden feststellen. Oftmals ist die Stichprobenvarianz größer als ihr Mittelwert (Overdispersion). Bei der Modellierung der Daten dieser Studie zeigte sich ebenfalls ein moderates Maß an Overdispersion. Ein bewährter Ansatz, um die vom Poisson-Modell unterschätze Varianz auszugleichen, ist das Adjustieren der Kovarianzmatrix der geschätzten Regressionskoeffizienten mittels des Quotienten aus Pearson Chi-Quadrat Goodness of Fit Statistik geteilt durch dessen zugehörige Freiheitsgrade [2].

4 Daten und Follow-Up

Die Berechnung familiärer Krebsrisiken erfolgte basierend auf Daten der schwedischen Familien-Krebs-Datenbank. Für die Studie wurden 8.148.737 Personen ausgewählt, die nach 1932 in Schweden geboren bzw. immigriert sind. Insgesamt wurden 355.166 medizinisch verifizierte Krebsfälle für diese Kohorte diagnostiziert. Um Inzidenzraten im Rahmen des Poisson Regressionsmodells vergleichen zu können, wurden Personenjahre und Krebsfälle während des Beobachtungszeitraums 1960-2010 gezählt. Die Verlaufsdaten wurden dabei entsprechend der zur Verfügung stehenden Informationen über Geschlecht, Altersgruppe, Dekade, sozioökonomischen Index und Wohnregion gruppiert. Die Daten wurden insbesondere entsprechend der familiären Vorgeschichte für verschiedene Krebsarten klassifiziert.

5 Programmbeispiel

Das eingangs beschriebene Poisson-Modell wird in folgendem SAS-Programm umgesetzt. Der Input-Datensatz enthält die oben beschriebenen Verlaufsdaten.

```
proc genmod data=input_followup order=data;
     class     fam_history (param=ref ref='Negative FH')
               sex agegroup socio_index county_index
               period_classes;
     model cases =  fam_history sex agegroup socio_index county_index
                    period_classes
          /dist=poisson link=log offset=log_pyrs scale=pearson;

     /* Berechnung bzw. Vergleich der relativen Risiken */
     estimate '1 Elternteil betroffen '     fam_history  1  0;
     estimate '1 Geschwister betroffen '    fam_history  0  1;
     estimate 'Test auf sign. Unterschiede' fam_history -1  1;

     /* Das folgende Statement legt einen neuen Datensatz RelRisk
          an, der die Ergebnisse der ESTIMATE-Funktion enthält */
     ods output Estimates=RelRisk;
run;
```

Wenn Interaktionen zwischen dem zu untersuchenden Risikofaktor und anderen Kovariaten in das Modell einbezogen werden sollen, muss der Interaktionsterm bei der ESTIMATE-Funktion bedacht werden. Interagieren beispielsweise `fam_history` und `sex`, dann wird der Term `fam_history*sex` in das MODEL-Statement aufgenommen. Da `sex` eine dichotome Variable ist (`sex`=0 für Frauen und `sex`=1 für Männer), lautet der ESTIMATE-Befehle zur Berechnung von geschlechtsspezifischen RRs wie folgt:

```
     /* RR wenn 1 Elternteil betroffen ist */
     estimate 'RR Söhne'      fam_history 1 0 fam_history*sex 1 0;
     estimate 'RR Töchter'    fam_history 1 0 fam_history*sex 0 0;

     /* RR wenn 1 Geschwister betroffen ist */
     estimate 'RR Bruder'     fam_history 0 1 fam_history*sex 0 1;
     estimate 'RR Schwester'  fam_history 0 1 fam_history*sex 0 0;
```

6 Ergebnisse und Interpretation

Für alle untersuchten Krebsarten waren Kinder bzw. Geschwister von Patienten einem signifikant erhöhtem Risiko ausgesetzt, an derselben Krebsart zu erkranken (siehe Tabelle 2). Während das familiäre Krebsrisiko im allgemeinen etwa doppelt so groß war, wie das von Personen ohne familiäre Vorbelastung, waren Risiken für manche Krebsarten bis zu 10-fach erhöht. Für nahezu jeden Krebs zeigten Geschwister ein höheres Risiko als Kinder von Patienten. Dies mag auf rezessive vererbte genetische Ursachen hindeuten oder auf schädliche Umwelteinflüsse während der Kindheit.

Tabelle 3: Familiäre Krebsrisiken, wenn ein Elternteil bzw. ein Geschwister betroffen ist

Krebsart	1 Elternteil betroffen			1 Geschwister betroffen			p-Wert für $RR_{Eltern} = RR_{Geschwister}$
	N[1]	RR	95% KI[2]	N	RR	95% KI	
Magen	172	1.78	(1.49-2.13)	46	2.97	(2.11-4.18)	**0.0090**[3]
Darm (Colon+Rectum)	2100	1.80	(1.71-1.90)	905	2.00	(1.84-2.17)	**0.0294**
Leber	101	1.78	(1.39-2.27)	30	2.07	(1.32-3.23)	0.56
Pankreas	140	2.09	(1.73-2.54)	54	2.73	(2.01-3.71)	0.15
Lunge	1108	1.92	(1.78-2.07)	712	2.50	(2.28-2.74)	**<.0001**
Brust	4989	1.78	(1.70-1.85)	3493	1.84	(1.75-1.94)	0.24
Eierstock	241	2.67	(2.32-3.07)	122	2.97	(2.44-3.61)	0.38
Prostata	4469	2.33	(2.24-2.43)	3390	2.59	(2.48-2.72)	**0.0005**
Hoden	32	3.96	(2.67-5.86)	82	6.94	(5.42-8.89)	**0.0173**
Niere	175	1.68	(1.40-2.01)	75	2.09	(1.59-2.75)	0.18
Harnblase	370	1.84	(1.61-2.12)	172	1.90	(1.55-2.32)	0.81
Melanom	643	2.55	(2.32-2.81)	584	2.74	(2.48-3.03)	0.31
Nervensystem	349	1.58	(1.38-1.80)	215	1.71	(1.44-2.03)	0.45
Schilddrüse	87	5.61	(4.25-7.40)	48	5.43	(3.74-7.87)	0.86
Non-Hodgkin-Lymphom	203	1.65	(1.41-1.93)	124	1.69	(1.38-2.06)	0.85
Hodgkin-Lymphom	18	2.59	(1.65-4.07)	40	9.60	(7.08-13.00)	**<.0001**
Leukämie	216	1.88	(1.61-2.20)	105	2.17	(1.73-2.71)	0.31

7 Schlussfolgerung

Die Implementierung eines Poisson Regressionsmodels in SAS ermöglicht die Berechnung von relativen familiären Krebsrisiken. Der GLM-Ansatz bietet dabei die Möglichkeit, sowohl den Einfluss des zu untersuchenden Risikofaktors zu bestimmen als auch weitere Confounder als Kovariaten miteinzubeziehen, um die Daten bestmöglichst zu modellieren. Mit Hilfe der PROC GENMOD Funktion in SAS lässt sich das Model einfach umsetzen und alle wichtigen Statistiken und Ausgabewerte können direkt berechnet werden. Die ODS OUTPUT Funktion bietet zusätzlich die Möglichkeit, die Ergebnisse in Datensätze zu speichern. Dies erweist sich insbesondere als nützlich, wenn die Daten zum Zwecke der Präsentation weiterverarbeitet werden sollen.

Literatur

[1] SAS Institute Inc. Sas/Stat 9.3 User's Guide. Cary, NC: SAS Institute Inc., 2011.
[2] Kianifard, F. and P. P. Gallo. "Poisson Regression Analysis in Clinical Research."

[1] Anzahl der Fälle mit positiver familiärer Vorgeschichte
[2] KI = Konfidenzintervall
[3] p-Werte, die signifikante Unterschiede zwischen den RRs zeigen, sind fettgedruckt hervorgehoben