Soz.- Präventivmed. 43 (1998) 312–321
0303-8408/98/060312-10 $ 1.50 + 0.20/0

Sozial- und Präventivmedizin

Willibald-Julius Stronegger[1], Andrea Berghold[2], Gilg U. H. Seeber[3]

[1] *Institut für Sozialmedizin, Universität Graz*
[2] *Institut für Medizinische Informatik, Statistik und Dokumentation, Universität Graz*
[3] *Institut für Statistik, Universität Innsbruck*

Epidemiologische und statistische Interaktionsmodelle und Folgen für die Regressionsanalyse

Zusammenfassung

Die unterschiedlichen Möglichkeiten, Interaktion zwischen zwei oder mehreren Expositionen in epidemiologischen Studien zu definieren, sowie die Unterscheidung von additiver und multiplikativer Wechselwirkung, führen immer wieder zu Missverständnissen über die Grundlagen solcher Definitionen. Statistische Regressionsmodelle für die Analyse von Ereignishäufigkeiten implizieren eine Definition von fehlender Interaktion, welche nicht immer eine für die vorliegenden Daten passende oder für die Fragestellung gewünschte ist. Die Autoren zeigen auf, dass unterschiedliche epidemiologische Kausalmodelle zu Interaktionsdefinitionen führen, die oftmals mit jenen der statistischen Modelle nicht übereinstimmen. Es zeigt sich, dass bei der Definition von Interaktion die Unterscheidung von Raten, Risiken und Odds zu berücksichtigen ist, sobald die Risiken nicht sehr klein sind. Anhand einer Datenanalyse wird illustriert, dass die üblicherweise vorausgesetzte multiplikative Interaktionsstruktur bei additiven Daten zu einem erheblichen Bias der Schätzwerte führt, sofern nicht alle signifikanten und nicht-signifikanten Interaktionsterme höherer Ordnung modelliert werden. Mittels nummerischer Verfahren berechnen die Autoren den bei der logistischen Regressionsanalyse entstehenden asymptotischen Bias ("Interaktions-Bias") für den Fall von zwei und drei dichotomen Expositionsvariablen.

In Regressionsmodellen liegt dann keine Interaktion zwischen zwei Prädiktoren vor, wenn der Effekt eines Prädiktors auf die Zielvariable nicht vom Wert des anderen Prädiktors abhängt, andernfalls spricht man von Interaktion oder Wechselwirkung. In der Epidemiologie ist die Zielvariable oft ein Ereignismass (Risiko, Rate oder Odds), und das Regressionsmodell wird in diesem Fall als *Ereignisdaten-Regressionsmodell* (event data regression) bezeichnet. In Ereignisdaten-Regressionsmodellen leitet sich die Definition von Interaktion zwischen zwei oder mehr unabhängigen Variablen bezüglich ihres Effektes auf ein Ereignismass aus den mathematischen Eigenschaften des Modells ab. Dies führt z.B. in der logistischen Regression oder der Cox-Regression zu einem multiplikativen Zusammenwirken der Effekte im Falle von fehlender Interaktion. Andererseits wurden in der Epidemiologie und Biologie eine Reihe formaler Kausalmodelle für den Zusammenhang zwischen Exposition und Erkrankungshäufigkeit entwickelt, aus welchen sich statistische Interaktionsbedingungen zwischen den Expositionen ableiten lassen. Diese stimmen nicht immer mit den von den Regressionsmodellen implizierten überein. Diese Kausalmodelle bezeichnen wir zur Abgrenzung von statistischen Modellbildungen und wegen ihres Bezuges zu epidemiologischen Anwendungen im folgenden als „Epidemiologische Interaktionsmodelle".

Zu den wichtigsten Kausalmodellen gehören das *Simple-Independent-Action-Model*[1] (SIAM), biologische Interaktionsmodelle (speziell das *Mehrstufenmodell*[2]), und das von Rothman[3] eingeführte *Sufficient-Component-Causes-Model* (SCCM). Das SIAM wurde entwickelt, um die gemeinsame Wirkung von zwei oder mehreren Expositionen richtig beschreiben beziehungsweise eine gegenseitige Verstärkung (Synergie) oder Abschwächung (Antagonie) prüfen zu können. Alle drei Modellklassen führen auf Definitionen

von Interaktion, welche substanzwissenschaftlichen Modellen besser entsprechen können als rein statistische Definitionen. Die Kausalmodelle ermöglichen dann eine verbesserte Vorhersage des Zusammenwirkens mehrerer Faktoren in zukünftigen Studien bzw. für neue Probanden. Neben der Vorhersage dienen diese Kausalmodelle auch der Überprüfung substanzwissenschaftlicher Modelle (z.B. der Karzinogenese) und damit zum besseren Verständnis der den Daten zugrundeliegenden Prozesse.

Im Folgenden geben wir zuerst ein Beispiel für nichtmultiplikative Interaktion, darauf zitieren wir statistische Interaktionsdefinitionen und stellen drei wichtige Kausalmodellklassen vor. Wir vergleichen die aus diesen Kausalmodellklassen folgenden Interaktionsbedingungen mit den Interaktionsdefinitionen der statistischen Regressionsmodelle und untersuchen die Fehler, welche bei der Analyse epidemiologischer Daten mittels logistischer Regression entstehen, wenn die vorliegende Interaktionsform nicht ausreichend modelliert wird. Für den Fall von zwei und drei dichotomen Expositionsvariablen berechnen wir mittels nummerischer Verfahren den asymptotischen Bias („Interaktions-Bias"), welchen die Schätzer der logistischen Regression bei additiven Interaktionen in den Daten aufweisen, wenn keine Interaktionsterme den Fehler korrigieren. Zuletzt werden Schlussfolgerungen für das praktische Vorgehen bei Regressionsanalysen diskutiert.

Analyse nichtmultiplikativer Ereignisdaten an einem Beispiel

Als illustratives Beispiel zur Einführung verwenden wir Daten aus einem im österreichischen Bundesland Steiermark durchgeführten Gesundheitssurvey. Es wurden Prädiktoren für das Vorliegen einer chronischen Atemwegserkrankung (kodiert als ja/nein) bei Männern gesucht. Wir fanden zwei statistisch signifikante Faktoren: Raucherstatus (Raucher/Nichtraucher) und einen Indikator für die Berufsschicht (manuelle vs. nichtmanuelle Tätigkeit). Wir adjustierten für Alter durch Stratifikation und analysierten die 1935 Männer in der Altersgruppe der 41–50-jährigen (Tabelle 1) mit einer verallgemeinerten logistischen Regression (s. Appendix A), welche sowohl die additive als auch die multiplikative Form fehlender Interaktion durch Vorgabe eines Parameters modellieren kann. Eine verallgemeinerte logistische Regression gehört zu den verallgemeinerten linearen Modellen (*generalized linear models*), deren Theorie den geeigneten Rahmen für unsere weiteren Analysen liefert. Tabelle 2 zeigt für jede durch Kombination der beiden Prädiktoren gebildete Untergruppe die absoluten („P_{ab}") sowie relativen („p_{ab}") Risiken für das Vorliegen einer Atemwegserkrankung.

Bei einem multiplikativen Zusammenwirken der Risiken wäre das gemeinsame relative Risiko für eine Atemwegserkrankung bei Rauchern mit manueller Tätigkeit vs. Nichtrauchern mit nichtmanueller Tätigkeit ungefähr gleich

$$p_{11} = p_{10} \cdot p_{01}$$
$$= 3{,}23 \cdot 4{,}13 = 13{,}34 \qquad (1)$$

und nicht der beobachtete Wert von 4,96 (Tabelle 2). Dies deutet auf das Vorhandensein einer Interaktion zwischen den beiden Faktoren hin, auch wenn sie trotz verhältnismässig grosser Fallzahl in der logistischen Regressionsanalyse nicht signifikant wurde (p = 0,24, vgl. Tabelle 3). Einer gängigen Praxis folgend wurde der nichtsignifikante Interaktionsterm in der Regressionsanalyse weggelassen. Die beiden geschätzten Haupteffekte liegen in ihren Werten zwischen den unterschiedlich grossen Effekten in den einzelnen Strata.

Zum Vergleich analysierten wir nun die Daten mit einem additiven

Häufigkeit	Berufsschicht: nichtmanuelle Tätigkeit		Berufsschicht: manuelle Tätigkeit		Total
chronische Atemwegserkrankung:	Nicht-Raucher	Raucher	Nicht-Raucher	Raucher	
ja (= 1)	3	3	58	44	108
nein (= 0)	187	43	1079	518	1827
Total	190	46	1137	562	1935

Tabelle 1. *Absolute Häufigkeit chronische Atemwegserkrankungen bei Männern (41–50 Jahre) nach Rauchstatus und Berufsschicht in Steiermark (n = 1935).*

Risiko (rel. Risiko)	Nichtmanuelle Tätigkeit	Manuelle Tätigkeit
Nichtraucher	$P_{00} = 0.0158$ ($p_{00} = 1$)	$P_{01} = 0.05$ ($p_{01} = 3.23$)
Raucher	$P_{10} = 0.0652$ ($p_{10} = 4.13$)	$P_{11} = 0.0783$ ($p_{11} = 4.96$)

Tabelle 2. *Risiko und relatives Risiko für chronische Atemwegserkrankungen bei Männern (41–50 Jahre) in Steiermark.*

Modell. Für ein additives Zusammenwirken beträgt das gemeinsame relative Risiko

$$p_{11} = p_{10} + p_{01} - 1 = 6{,}36 \qquad (2)$$

welches mit dem beobachteten Wert von 4,96 besser übereinstimmt als Gleichung (1). Die Abweichung lässt auf ein leicht unteradditives Zusammenwirken schliessen. Der Interaktionsterm ist wieder nicht signifikant, die Haupteffekte (in beiden Fällen als Odds Ratio ausgedrückt) der Regressionsanalyse sind nun deutlich grösser und beide signifikant (Tabelle 3).

Vergleichen wir die Resultate der multiplikativen und der additiven Regression (Tabelle 3), sehen wir deutliche Unterschiede bei den geschätzten Haupteffekten, sowohl in der Grösse als auch in der Signifikanz (d.h. das 95%-Konfidenzintervall enthält nicht die 1). Die Verbesserung der Modellanpassung zum saturierten Modell gemessen in Devianz ist 1,4 bzw. 0,3 auf einen Freiheitsgrad im multiplikativen bzw. additiven Modell. Empirisch kann deshalb nicht ohne weitere Information entschieden werden, welches Modell die Daten „richtiger" beschreibt und welche Schätzwerte daher die besseren sind. Dass in solchen Situationen standardmässig mit dem multiplikativen Modell (d.h. der logistischen Regression) und nicht dem additiven gearbeitet wird, ist nicht immer begründet und von der Datenlage her gesehen sinnvoll, sondern darauf zurückzuführen, dass das multiplikative Modell (1) die angenehmeren mathematischen Eigenschaften aufweist.

Zu bedenken ist, dass das multiplikative Zusammenwirken von Faktoren nur einen Punkt auf einem Kontinuum von möglichen Interaktionsdefinitionen bildet. Die Einbeziehung weiterer Kriterien für die Auswahl der Interaktionsform erscheint uns hier notwendig zu sein. Ein möglicher Weg ist die Zugrundelegung theoretischer Modelle wie es epidemiologische Kausalmodelle sind, ein anderer Weg ist die Anpassung der Interaktionsstruktur an die Daten mittels Schätzung eines zusätzlichen „Interaktionsparameters" im Modell. Dies kann durch Parametrisierung der Linkfunktion in Ereignisdaten-Regressionsmodellen geschehen. Für diesen Zweck wurden para-

Prädiktoren	Maximum Likelihood Schätzwerte (Odds Ratio mit 95%-Konfidenzintervall)			Schätzwerte stratifizierte Analyse[b] (Odds Ratio)
	Logistische Regression ($a = 0$)[a]	Additive Regression ($a = 1$)[a]	Regression mit geschätztem Link $a = 2{,}15$[a] (unteradditiv)	
Beruflicher Status	2,27 (0,98–5,25)	**3,05** (1,75–4,35)	**3,35** (2,21–4,16)	3,35
Raucher	**1,67** (1,13–2,48)	**2,99** (1,39–4,58)	4,35 (0,70–5,97)	4,35
Interaktionsterm	weggelassen: nicht signifikant (p = 0,24)	weggelassen: nicht signifikant	kein freier Parameter mehr vorhanden	–
Devianz (Freiheitsgrade)	821,2 (1932)	820,1 (1932)	819,8 (1931)	–

[a] a bezeichnet den Linkparameter der Linkfamilie nach Guerrero-Johnson.
[b] Analyse mit Status = 0 und Raucher = 0 als Referenzkategorie.

Tabelle 3. *Resultate der logistischen, additiven und unteradditiven verallgemeinerten Regression mit parametrisierter Linkfunktion (n = 1935).*

metrisierte Linkfamilien mit unterschiedlichen mathematischen Eigenschaften vorgeschlagen[4], wobei die Familie nach Guerrero-Johnson[5] nach unseren Untersuchungen die besten numerischen Eigenschaften aufweist und deshalb dem vorgeschlagenen Regressionsmodell zugrundeliegt. Eine optimale Anpassung der in dieser Weise verallgemeinerten Regression auf ein interaktionsfreies Modell mittels Schätzung des Parameters *a* der parametrisierten Linkfamilie nach Guerrero-Johnson[5] führt bei unseren Daten wie erwartet auf ein unteradditives Modell. Dieses liefert wiederum andere Schätzwerte und Signifikanzen für die Haupteffekte als das multiplikative oder additive Modell, die jedoch nun mit jenen der stratifizierten Analyse übereinstimmen (Tabelle 3).

Interaktion in statistischen Modellen

Masszahlen, die das Eintreten von Ereignissen quantifizieren, werden in der Epidemiologie als *„Ereignismasse"* bezeichnet. Diese können gemäss ihren mathematisch-statistischen Eigenschaften in Risiken, Raten und Odds eingeteilt werden. Entsprechend ihrer jeweiligen epidemiologischen Anwendung beschreiben Risiken z.B. Prävalenzen und Inzidenzproportionen (früher irreführend auch bezeichnet als: kumulative Inzidenzen), Raten vor allem Hazardraten (auch: Inzidenzdichten), und Odds beschreiben z.B. Prävalenzodds oder in approximativer Weise Risiken. Odds werden von der logistischen Regression modelliert, Raten von der Poisson- und Proportional-Hazard-Regression. Risiken werden zumeist nicht direkt, sondern approximativ mittels eines der genannten Regressionsmodelle modelliert, mit der Begründung, dass Odds und Raten eine gute Approximation im Falle kleiner Risiken (etwa $<10\%$) sind. In vielen Anwendungen, etwa bei der Analyse von Verhaltensweisen (Rauchen etc.) oder Beschwerden, trifft dies jedoch nicht zu, und die Verwendung von *„Effektmassen"* wie Odds Ratio oder Rate Ratio zur Darstellung der Resultate ist fragwürdig.

Zur Beschreibung der Bedingungen für fehlende Interaktion in Regressionsmodellen beschränken wir uns auf den Fall zweier dichotomer unabhängiger Variablen A und B, welche jeweils die Werte 0 („nicht exponiert") und 1 („exponiert") annehmen können. Für jede Ausprägung a bzw. b von A bzw. B beschreibe dann E_{ab} das zugehörige Ereignismass und $e_{ab} := E_{ab}/E_{00}$ das zugehörige *(relative) Effektmass*. In der logistischen Regression folgt für die Odds $O_{ab} = P_{ab}/(1-P_{ab})$ bei fehlender Interaktion zwischen A und B aus der Modellgleichung:

$$O_{11} \cdot O_{00} = O_{10} \cdot O_{01} \qquad (3)$$

und für das Odds Ratio $o_{ab} = O_{ab}/O_{00}$ folgt:

$$o_{11} = o_{10} \cdot o_{01} \qquad (4)$$

also eine *Multiplikation* der Odds und Odds Ratios. Die Relationen (3) und (4) entsprechen für kleine Risiken $P = O/(1+O) < 0{,}1$ approximativ ebenfalls einer Multiplikation der Risiken und relativen Risiken, für grosse Risiken $P > 0{,}9$ entsprechen sie annäherungsweise einer *additiven* Relation der Risiken:

$$P_{11} + P_{00} = P_{10} + P_{01} \qquad (5)$$

und der relativen Risiken:

$$p_{11} = p_{10} + p_{01} - 1 \qquad (6)$$

Die gleichen multiplikativen Relationen (3) und (4) gelten bei Poisson- und Proportional-Hazard Regression nicht für die Odds, sondern für die Raten R_{ab} bzw. relativen Raten (Rate Ratios) $r_{ab} = R_{ab}/R_{00}$:

$$R_{11} \cdot R_{00} = R_{10} \cdot R_{01} \text{ bzw.} \qquad (7)$$

$$r_{11} = r_{10} \cdot r_0 \qquad (8)$$

Der Zusammenhang zwischen Risiko und Rate ist nicht wie jener zwischen Risiko und Odds eindeutig bestimmt, sondern abhängig vom Studiendesign[6], sodass über die implizierten Interaktionsbedingungen auf den Risiken bzw. Odds keine generelle Aussage möglich ist. Bei sehr kleinen Raten ($<0{,}001$) unterscheiden sich in der Regel Raten, Risiken und Odds kaum und die Interaktionsdefinition ist daher auf jedem Ereignismass gleichwertig.

Bei Querschnittstudien können aus den Daten nur Prävalenzen P (d.h. Risiken) bzw. Prävalenzodds PO (d.h. Odds) berechnet werden. Wie sie sich bezüglich Interaktion verhalten, hängt einerseits vom Interaktionsverhalten der zugrundeliegenden Raten R und andererseits vom Zusammenhang zwischen Rate und Prävalenz ab. Nach Kleinbaum et al.[7] müssen zwei prinzipielle Typen von Querschnittstudien unterschieden werden. Bei der „density-type" Querschnittstudie (dynamische, aber stationäre Kohorte) ist die Prävalenzodds PO das Produkt aus mittlerer Erkrankungsdauer D und Hazard-Rate (Inzidenzdichte):

$$PO := P/(1-P) = R \cdot D \qquad (9)$$

d.h. bei additiv bzw. multiplikativ zusammenwirkenden Raten kombinieren auch die Odds additiv bzw. multiplikativ. Der Einsatz der logistischen Regression bei multiplikativ kombinierenden Raten ist hier sinnvoll. Bei der „cumulative-type" Querschnittstudie, bei der eine fixe Population über einen Zeitraum der Dauer T „at risk" zur Krankheitsentstehung ist, gilt für den Zusammenhang zwischen Prävalenz P zum Zeit-

punkt T und Krankheitsinzidenzdichte R:

$$-\log(1-P) = R \cdot T \qquad (10)$$

d.h. die komplementären Logarithmen der Risiken verhalten sich wie die Raten. Hier erscheint die Verwendung der Binomialregression mit komplementär-log-log Linkfunktion bei multiplikativ kombinierenden Raten sinnvoll, was auch vor kurzem für diesen Studientyp diskutiert wurde[8]. Bei Fall-Kontroll-Studien ist die Situation je nach der Studiendesignvariante unterschiedlich[7]. Aus den Daten lässt sich grundsätzlich kein Erkrankungsrisiko, sondern nur das Expositions-Odds Ratio (EOR) direkt berechnen. Das EOR kann sich je nach Designvariante verhalten wie das Risiko-Odds Ratio, wie das Rate Ratio oder wie das Risk Ratio, und je nachdem können andere Interaktionsbedingungen gelten. Bei Kohortenstudien kann die Rate R durch Proportional-Hazard-Regressionen (Cox-Regression) direkt modelliert werden, was ein multiplikatives Zusammenwirken der Raten impliziert.

Interaktion in epidemiologischen Modellen

Simple-Independent-Action-Model

Zu den bedeutendsten Interaktionsmodellen zählt das „Simple-Independent-Action-Model" (SIAM), das in verschiedenen Kontexten als „Disease-Hit Model", „Single-Hit Model" und „Vulnerability Model" eingeführt wurde[1] (s. Appendix B). In seiner Grundform beschreibt es mengentheoretische Relationen zwischen Risiken für das Eintreten von Ereignissen. Rothman[9] erweiterte dieses Modell auf die Situation mit Hintergrundfaktoren C (d.i. mit Grundrisiko P_{00}) und erhält für das interaktionsfreie Zusammenwirken zweier binärer Faktoren A und B die Bedingung:

$$\begin{aligned} P_{11} - P_{00} = {} & (P_{10} - P_{00}) \\ & + (P_{01} - P_{00}) - (P_{10} - P_{00}) \\ & \cdot (P_{01} - P_{00})/(1 - P_{00}) \end{aligned} \qquad (11)$$

Die Erhöhung des gemeinsamen Risikos P_{11} gegenüber dem Grundrisiko P_{00} ist also nach Gleichung (11) gleich der Summe der Risikoerhöhungen für die einzelnen Faktoren abzüglich eines Termes, der umso grösser ist, je grösser das Grundrisiko ist. Wir haben nach diesem Modell folglich sowohl für Risiken wie für relative Risiken Sub-Additivität bei Nicht-Interaktion. Walter und Holford[10] zeigten mittels ihres „Single-Hit Model", welches annimmt, dass den Risiken im SIAM stationäre Poissonprozesse mit Raten R_A, R_B und R_C zugrundeliegen, dass die Raten dann eine exakte additive Relation (12) erfüllen:

$$\begin{aligned} R_{11} := {} & (R_A + R_C) \\ & + (R_B + R_C) - R_C \\ = {} & R_{10} + R_{01} - R_{00} \end{aligned} \qquad (12)$$

Mehrstufenmodelle

Es gibt keinen generellen Konsens, was unter biologischer Interaktion zu verstehen sei[11,12]. Eine Definition der Interaktion von allgemeinerer Gültigkeit ist auf dieser Basis also nur schwer zu finden. Als Beispiel führen wir das Initiation-Promotion Modell an: Wenn Faktor A die Anzahl suszeptibler Zellen erhöht und Faktor B die Transformation dieser Zellen in maligne fördert, besteht kein Konsens, ob diese Faktoren unabhängig oder synergistisch wirken.

Etwas anders ist die Situation, wenn man sich nur mit einer ganz bestimmten Pathogenese beschäftigt. Die Übereinstimmung bezüglich einer genauen Definition biologischer Interaktion dürfte auch hier nur schwer zu erreichen sein, ist aber auch nicht notwendig, da bei einem bekannten Wirkmechanismus auch ohne eine solche Terminologie die Folgen für das Krankheitsgeschehen abgeleitet werden können[12].

Wir beschränken uns deshalb auf eine interessante Kategorie quantitativer biologischer Kausalmodelle, das *Mehrstufenmodell* der Karzinogenese („*Multi-Stage-Model*"), das sich empirisch gut bewährt hat und für eine grosse Gruppe chronischer Krankheiten geeignet erscheint. *Mehrstufenmodelle* der Karzinogenese gehen auf früheste quantitative Theorien der Karzinogenese zurück, die einen Übergang von einer gesunden zu einer bösartigen Zelle modellieren. Es wurden im Laufe der Zeit verschiedene Mehrstufenmodelle der Krebsentstehung vorgeschlagen[2], wobei wir uns auf das *Armitage-Doll-Modell*[13] konzentrieren, welches experimentell einige Bestätigung erfahren hat. Es nimmt an, dass Krebs aus einer einzelnen, anfangs normalen Zelle, entsteht, welche nach und nach eine Reihe von Transformationen durchläuft, bis sie zu einer zur malignen Entartung fähigen Zelle wird.

Es kann gezeigt werden, dass auf derselben Transformationsstufe wirkende Kausalfaktoren zu *additiven* Inzidenzraten führen und Faktoren, welche auf unterschiedlichen Stufen ansetzen, *multiplikativ* wirken[14]. Komplizierter wird das Bild, wenn Faktoren zugleich mehrere Mutationsraten beeinflussen. So z.B. wenn Faktor A auf die i-te und j-te Rate wirkt, B nur auf die j-te. Die Inzidenzrate liegt dann zwischen Additivität und Multiplikativität. Es ist leicht einzusehen, dass die Interaktion auch bei noch komplexeren Wirkmustern immer zwischen Additivität und Multiplikativität liegt, sofern die Interaktion auf den einzelnen Mutationsraten additiv bleibt[15]. Es gibt eine Reihe empirischer Resultate, die eine untermultiplikative bzw. additive Interaktion bestätigen. So

zeigen Lund et al.[16], dass die bekannten Risikofaktoren beim postmenopausalen Brustkrebs additiv und nicht multiplikativ zusammenwirken.

Sufficient-Component-Causes-Model

Das von Rothman[3] eingeführte *Sufficient-Component-Causes-Model* (SCCM) ist eines der wenigen konzeptionellen Kausalmodelle der Epidemiologie. Es unterscheidet zwischen *Suffizienten-Ursachen*, welche jede für sich hinreichend für die Krankheitsentstehung ist, und den *Komponenten-Ursachen*, welche Teil mindestens einer Suffizienten-Ursache sind und für sich allein nicht ausreichen, die Krankheit zu verursachen. Die Suffizienten-Ursachen bestehen also selbst aus Komponenten-Ursachen. So einfach dieses Modell ist, so wenig kann es durch die bestehenden statistischen Modelle adäquat beschrieben werden. Koopman[17,18] greift das SCCM wieder auf und folgert aus dem SCCM die Additivität der Raten bei Nicht-Interaktion. Wir gehen auf dieses Modell nicht weiter ein, merken aber an, dass es nach den Untersuchungen von Koopman mehr die additive als die multiplikative Synergie als Form von Nicht-Interaktion nahelegt.

„Interaktions-Bias" bei der Analyse nichtmultiplikativer Daten ohne Interaktionsterme

Wir nehmen nun an, dass uns Ereignisdaten mit einem additiven Zusammenwirken der Odds vorliegen. Werden diese Daten mit einem multiplikativen Standardverfahren wie der logistischen Regression ohne Interaktionsterme analysiert, z.B. weil diese nicht signifikant werden oder ihre Anzahl zu gross wäre (Verlust an Power), so entspricht dies einer Fehlspezifikation der Linkfunktion, also des systematischen Teils eines verallgemeinerten linearen Regressionsmodells (*generalized linear models*), da dann die durch die Linkfunktion vorgenommene Transformation der Daten auf kein additives Zusammenwirken der Prädiktoren entsprechend der Prädiktorgleichung $y := x_1\beta_1 + \ldots + x_d\beta_d$ führt. Es ist bekannt[19], dass dann zusätzlich zum bei endlichen Stichproben immer vorhandenen Bias des Maximum-Likelihood-Schätzers ein asymptotischer Bias auftritt. Wir bezeichnen diesen asymptotischen Bias als „*Interaktions-Bias*", da er bei nicht ausreichender Modellierung der vorhan-

Wahre Odds Ratios		Geschätzte Odds Ratios (bei $P_0 = 0{,}01$)		Geschätzte Odds Ratios (bei $P_0 = 0{,}50$)	
OR_1	OR_2	or_1	or_2	or_1	or_2
1,5	2	1,33	1,80	1,38	1,83
1,5	3	1,25	2,60	1,35	2,66
2	2	1,67	1,67	1,75	1,75
2	5	1,34	3,67	1,67	3,97
3	3	2,01	2,01	2,35	2,35
3	5	1,68	3,02	2,30	3,65

Tabelle 4. *Interaktions-Bias bei additiven Daten mit 2 dichotomen Prädiktoren (jede Kombination tritt mit Wahrscheinlichkeit $P_{ij} = 0{,}25$ auf): $F(y) = (e^y)/(1 + e^y)$ inverse logistische Linkfunktion.*

Wahre Odds Ratios (additive Odds)			Geschätzte Odds Ratios (mit logistischer Regression)		
OR_1	OR_2	OR_3	or_1	or_2	or_3
1,5	1,5	1,5	1,33	1,33	1,33
1,5	1,5	3	1,22	1,22	2,33
1,5	1,5	5	1,16	1,16	3,67
1,5	2	3	1,20	1,45	2,15
1,5	3	3	1,17	1,90	1,90
1,5	3	5	1,13	1,63	2,79
2	2	2	1,50	1,50	1,50
2	2	5	1,29	1,29	3,00
2	4	5	1,23	1,88	2,35
3	3	3	1,68	1,68	1,68
2	3	6	1,23	1,52	3,02
2	3	0,166	1,64	2,86	0,66
2	0,5	0,2	3,89	0,54	0,36
1,5	0,75	0,25	2,01	0,71	0,33

Tabelle 5. *Interaktions-Bias bei additiven Daten mit 3 dichotomen Prädiktoren (jede Kombination tritt mit Wahrscheinlichkeit $P_{ijk} = 0{,}125$ auf): $P_0 = 0{,}01$; $F(y) = (e^y)/(1 + e^y)$ inverse logistische Linkfunktion.*

denen (auch nichtsignifikanten) Interaktionen zwischen den unabhängigen Variablen auftritt.
Czado & Santner[19] geben ein Integralgleichungssystem an, das bei Angabe der „wahren" Responsefunktion H(y), d.h. der die Daten beschreibenden inversen Linkfunktion, und der in der Datenanalyse angenommenen Responsefunktion F(y) die Berechnung des asymptotischen Bias erlaubt. Da wir im weiteren immer nur mit der Responsefunktion, welche die Umkehrfunktion der Linkfunktion ist, arbeiten, bezeichnen wir sie einfach als „Linkfunktion".
Wir lösten mittels nummerischer Verfahren das Integralgleichungssystem für ein logistisches Regressionsmodell mit zwei und drei dichotomen unabhängigen Variablen und dem logistischen Link $F(y) = (e^y)/(1 + e^y)$ als angenommener Linkfunktion und dem additiven Link $H(y) = (1+y)/(2+y)$ (Herleitung s. Appendix C) als „wahrer" Linkfunktion. Für „kausale" Faktoren (d.h. Odds Ratio >1) führt der asymptotische Bias zu einer deutlichen Unterschätzung der tatsächlichen Haupteffekte (Tabelle 4 und 5), und dieser Bias nimmt mit der Effektgrösse, der Anzahl der Prädiktoren und der Abweichung des Grundrisikos P_0 von 0,5 zu. So werden z.B. bei drei Variablen die in den Daten vorliegenden Odds-Ratios mit Werten von 2, 2 und 5 durch die logistische Regression zu 1,29, 1,29 und 3,00 geschätzt (Tabelle 5). Simulationsstudien zeigten, dass der Bias für endliche Stichproben in der Regel noch grösser als der asymptotische ist.

Diskussion: Schlussfolgerungen für die Regressionsanalyse

Die aus epidemiologischen und biologischen Kausalmodellen folgenden Bedingungen für fehlende Interaktion stimmen in den meisten Fällen nicht mit den in den statistischen Modellen vorausgesetzten überein und führen zu einem nichtmultiplikativen Zusammenwirken, vor allem der Raten und damit meist auch der Risiken[6]. Es zeigt sich, dass die Analyse nichtmultiplikativer Risiken mittels multiplikativer Modelle mit einem zum Teil beträchtlichen asymptotischen Bias („Interaktions-Bias") bei den Haupteffektschätzwerten verbunden ist, wenn irrtümlich signifikante aber auch nichtsignifikante Interaktionsterme weggelassen werden. Die korrigierenden Interaktionsterme werden selbst bei grossen Fallzahlen selten signifikant und deshalb oft fälschlicherweise weggelassen und ihre Anzahl wird bereits bei wenigen unabhängigen Faktoren zu gross, um sie alle berücksichtigen zu können.
Eine wesentliche Folgerung aus unserer Arbeit besteht darin, dass bei epidemiologischen Daten nicht nur aufgrund empirischer Resultate[16], sondern auch aufgrund konzeptueller Überlegungen mit nichtmultiplikativer Interaktion zu rechnen ist, die Multiplikativität sich sogar entgegen der üblichen Analysepraxis als ein beliebiger Spezialfall unter anderen erweist und darum überprüft werden muss. Das den epidemiologischen Daten entsprechende Kausalmodell sollte nach Möglichkeit bei der Datenanalyse und der statistischen Modellbildung berücksichtigt werden, um eine mögliche Fehlspezifikation des Modells auszuschliessen. Ist aus theoretischen Überlegungen über die Interaktionsform nichts bekannt, so sind parametrische Regressionsmodelle, welche auch nichtmultiplikative Nicht-Interaktion beschreiben, unter Umständen geeignete Verfahren, um die durch *Interaktions-Bias* stark verzerrten Haupteffektschätzwerte der Standardmodelle zu vermeiden. Diese arbeiten mit parametrisierten Linkfamilien, welche für die Erweiterung der Ereignisdaten-Regressionsmodelle auf allgemeinere Interaktionsformen geeignet sind[20]. Wir beschränkten uns auf dichotome Prädiktorvariablen, die Ereignisdaten-Regressionsmodelle sind aber in der gleichen Form auch zur Analyse quantitativer Prädiktoren geeignet.
Bei der Modellwahl in Regressionsanalysen sollte grundsätzlich unterschieden werden, ob das Ziel in der optimalen Prädiktion der Zielvariablen bei neuen Probanden besteht, oder das Regressionsmodell zur Erklärung des Zusammenhangs zwischen Zielgrösse und potentiellen Prädiktoren dient. Die erste Aufgabenstellung ist vor allem in der praktischen Anwendung der Modelle von Bedeutung, die zweite entspricht dem Erkenntnisinteresse der wissenschaftlichen Forschung. In der epidemiologischen Literatur[21] wird betont, dass die beiden Zielsetzungen unterschiedliche Modellselektionsstrategien erfordern. Die Prädiktionsaufgabe erfüllt jenes Modell am besten, welches die Zielvariable bei möglichst geringer Komplexität am besten vorhersagt, unabhängig von den dabei jeweils einbezogenen Variablen. Automatische Variablenselektionsverfahren sind hier sinnvolle Strategien. Für diese Zielsetzung schlagen wir bei verallgemeinerten Regressionsmodellen vor, den Linkparameter *a* zu schätzen und den Schätzwert zur Prädiktion zu verwenden.
Für die zweite Aufgabenstellung sind automatische Verfahren nicht geeignet[21], da ein mit wissenschaftlichen Theorien in Einklang stehendes Modell gefunden werden soll, das auch eine aussagefähige Interpretation der Resultate zulässt. Bei der verallgemeinerten Regression ist es dann zweckmässig zu entscheiden, ob eines von zwei oder mehreren theoretisch möglichen Modellen (z.B. ein multiplikatives oder additives) besser den Daten entspricht und die endgültige Analyse mit diesem durchzuführen. Dies ermöglicht dann eine Aussage über die Interaktionsform und gewährleistet die Inter-

pretierbarkeit der Regressionskoeffizienten, welche bei „Zwischenmodellen" nicht gegeben ist. Modelle mit Interaktionstermen ebenso wie nichtmultiplikative und nichtadditive Modelle führen bei der Beschreibung der Effekte zu Interpretationsproblemen, weshalb parametrisierte Modelle hier vor allem für eine Entscheidung zwischen dem additiven und dem multiplikativen Modell eingesetzt werden sollten.

Summary

Epidemiological and statistical models of interaction and their analysis by regression models

Different ways to define interaction between exposition factors in epidemiological studies as well as the choice between additive and multiplicative no interaction leads frequently to confusion during data analysis. In their standard form methods of event data analysis such as Poisson or logistic regression assume a multiplicative parameterization of no interaction. However, evidence from empirical investigations as well as causal models of disease etiology, e.g. the simple independent action model of Finney or the sufficient-component-causes model of Rothman, suggest additive or other kinds of non-multiplicative concepts of no interaction. For additive structured data we illustrate the asymptotic bias ("interaction-bias") of main effect estimates which are based on inappropriate data analysis using multiplicative models and omitting significant or non-significant interaction terms. We show that both the epidemiological study design as well as the underlying causal model are determinants of the interaction structure of the data and should be considered in the model selection process. The definition of interaction should distinguish between risk, rate and odds if risks are not very small. Using generalized linear models with parametrical link functions we are able to analyze non-multiplicative interaction structures.

Résumé

La définition des interactions épidémiologique et statistique et leur analyse à l'aide de modèles de régression

Les diverses possibilités de définir dans des études épidémiologiques l'interaction survenant entre deux ou plusieurs expositions ainsi que de faire la différence entre l'interaction additive et multiplicative prêtent à confusion en ce qui concerne le fondement de telles définitions. Les modèles de régression statistiques employés dans l'analyse de fréquences d'événement impliquent une définition, sans qu'il y ait interaction, définition qui n'est pas toujours appropriée dans le contexte des données en question ni désirée dans le cadre du thème posé. A l'aide d'une analyse de données, les auteurs visent à illustrer que la structure d'interaction multiplicative habituellement présumée pour des données additives mène à un biais considérable dans les valeurs estimées, à défaut d'un modèle intégrant tous les termes d'interaction significatifs et non-significatifs. Les auteurs emploient des procédures numériques pour illustrer le biais asymptotique («biais d'interaction») qui survient dans l'emploi d'analyses de régression logistique, à l'exemple de deux et de trois variables d'exposition dichotomes. Les auteurs démontrent que de différents modèles épidémiologiques causals provoquent des définitions d'interaction qui contredisent souvent ceux des modèles statistiques. Il est donc évident que la définition d'interactions exige de tenir compte des différences entre les ratios, les risques et les odds une fois que les risques dépassent une certaine limite.

Literaturverzeichnis

1 *Finney DJ.* Quantal responses to mixtures. In: Probit Analysis. 3rd ed. Cambridge: Cambridge University Press, 1971: 231–283.

2 *Breslow NE, Day NE.* Statistical Methods in Cancer Research: Vol. II – The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer, 1987.

3 *Rothman KJ.* Causes. Am J Epidemiol 1976; *104*: 587–592.

4 *Moolgavkar SH, Venzon DJ.* General relative risk regression models for epidemiologic studies. Am J Epidemiol 1987; *126*: 949–961.

5 *Guerrero VM, Johnson RA.* Use of the Box-Cox transformation with binary response models. Biometrika 1982; *69*: 309–314.

6 *Stronegger W-J, Berghold A.* Re: „Biologic Synergism and Parallelism". Am J Epidemiol 1998; *147*: 89.

7 *Kleinbaum DG, Kupper LL, Morgenstern H.* Epidemiologic research: Principles and quantitative methods. New York: Van Nostrand Reinhold, 1982.

8 *Martuzzi M, Elliott P.* Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. Ann Epidemiol 1998; *8*: 52–55.

9 *Rothman KJ.* The estimation of synergy or antagonism. Am J Epidemiol 1976; *103*: 506–511.

10 *Walter SD, Holford TR.* Additive, multiplicative, and other models for disease risks. Am J Epidemiol 1978; *108*:341–346.
11 *Weinberg CR.* Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. Am J Epidemiol 1986; *123*: 162–173.
12 *Rothman KJ, Greenland S, Walker AM.* Concepts of interaction. Am J Epidemiol 1980; *112*:467–470.
13 *Armitage P, Doll R.* The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954; *8*:1–12.
14 *Siemiatycki J, Thomas DC.* Biological models and statistical interactions: An example from multistage carcinogenesis. Int J Epidemiol 1981; *10*:383–387.
15 *Kaldor JM, L'Abbé KA.* Interaction between human cacinogens. In: Complex Mixtures and Cancer Risk: IARC Scientific Publications No. 104. Lyon: International Agency for Research on Cancer, 1990: 35–43.
16 *Lund E.* Comparison of additive and multiplicative models for reproductive risk factors and postmenopausal breast cancer. Stat Med 1995; *14*:267–274.
17 *Koopman JS.* Causal models and sources of interaction. Am J Epidemiol 1977; *106*:439–444.
18 *Koopman JS.* Interaction between discrete causes. Am J Epidemiol 1981; *113*:716–724.
19 *Czado C, Santner TJ.* The effect of link misspecification on binary regression inference. J Stat Plan Inference 1992; *33*:213–231.
20 *Stronegger W-J, Seeber GUH.* Modelling of Non-multiplicative Interactions for Epidemiologic Causal Models Using Families of Link Functions. In: Minder C.E., Friedl H., eds. Good statistical practice. Wien: Österreichische Statistische Gesellschaft, 1997: 280–284. (Seeber G.U.H., ed. Schriftenreihe der Österreichischen Statistischen Gesellschaft; Bd. 5)
21 *Clayton D, Hills M.* Statistical Models in Epidemiology. Oxford: Oxford University Press, 1993: 271–281.
22 *Francis B, Green M, Payne C,* et al. The GLIM System. Release 4 Manual. Oxford: Claredon Press, Reprint, 1994.

Korrespondenzadresse

Mag. Dr. W.-J. Stronegger
Institute of Social Medicine
Universitätsstrasse 6/I
A-8010 Graz
Fax +43 316 380 9665
willibald.stronegger@kfunigraz.ac.at

Appendix

(A) In der *verallgemeinerten logistischen Regression* mit Linkfamilie nach Guerrero-Johnson[5] lautet die Regressionsgleichung für das Risiko P in Abhängigkeit von den d Prädiktoren x_1 bis x_d:

$$P/(1-P) = \text{Odds}(x_1, \ldots, x_d) = [1 + a(\beta_0 + x_1\beta_1 + \ldots + x_d\beta_d)]^{1/a}$$

In dieser Gleichung bezeichnet *a* den Linkparameter, der für das multiplikative Standardmodell gegen 0 konvergiert und für das additive Modell gleich 1 ist. Linkparameterwerte kleiner als 0 entsprechen einem über-multiplikativen, Werte grösser als 1 einem unter-additiven Modell.
Für die Berechnung verallgemeinerter event-data Regressionsmodelle mit festem Linkparameter *a* eignet sich besonders das Softwarepaket für *generalized linear models*, GLIM4[22], welches die Definition parametrisierter Linkfunktionen erlaubt. Die Schätzung des Linkparameters *a* erfordert zusätzlich einen speziellen Schätzalgorithmus, der von den Autoren als GLIM-Macro erhältlich ist.
(B) Das Modell der *„Simple Independent Action"* (SIAM) geht zurück auf Überlegungen bezüglich der Toxizität von Giften, die gemeinsam verabreicht werden. Dabei entstanden unmittelbar Fragen darüber, wie die gemeinsame Wirkung von Substanzen zu beschreiben und klassifizieren sei. Man kann das SIAM nicht nur mengentheoretisch, sondern auch direkt über Response-Raten anschaulich definieren. Zur Definition wird die Proportion getöteter Tiere bei Giftgabe verwendet, also ein probabilistischer Begriff, und der Begriff der *Suszeptibilität* gegenüber zweier Gifte a oder b, d.h. das Reagieren auf die Giftgabe a oder b durch Tod. Bezeichne nun A bzw. B das Ereignis auf das Gift a bzw. b suszeptibel zu sein. Wird Gift a bzw. b allein verabreicht, erhalte man die Response-Raten P_A bzw. P_B. Bei Verabreichung beider Gifte und *stochastisch unabhängiger* Suszeptibilität wird man unter den $1-P_A$ Non-Respondern bezüglich Gift a noch einen Anteil von P_B Respondern bezüglich Gift b haben, insgesamt also:

$$P_{A\&B} = P_A + P_B(1-P_A) = 1-(1-P_A)(1-P_B)$$

Rothman[9] erweitert dieses probabilistische Modell um Hintergrundursachen C und erhält Gleichung (11). Ein naheliegender Schritt, da es auch ohne Vorliegen spezifischer Ursachen immer ein bestimmtes Grundrisiko – das von Hintergrundursachen bewirkt angenommen wird – zu erkranken gibt.

(C) Das Risiko P wird in der logistischen Regression mittels einer Binomialverteilung beschrieben und nach einer Transformation $\log[P/(1-P)] = y$ durch eine lineare Prädiktorgleichung $y := x_1\beta_1 + \ldots + x_d\beta_d$ modelliert. Die Auflösung dieser Gleichung nach dem Risiko P ergibt die sogenannte Responsefunktion bzw. inverse Linkfunktion $H_{mult}(y) = P = (e^y)/(1+e^y)$ der logistischen Regression. Im Falle additiver Odds lautet die Regressionsgleichung $P/(1-P) = 1+y$, d.h. die Odds werden nur linear und nicht logarithmisch transformiert. Mit $1+y$ anstatt nur mit y wird hier aus Symmetriegründen gearbeitet. Durch Auflösen nach P folgt die additive inverse Linkfunktion zu $H_{add}(y) = (1+y)/(2+y)$.