Die Ermittlung von Risikofaktoren: Das Verfahren der logistischen Regression

Sieber, M

Abstract: Zusammenfassung: Die logistische Regression ermöglicht die Bestimmung von Risikofaktoren für ein Zielkriterium, wobei das Zielkriterium zweistufig ist, z.B. krank/gesund. Aus den Regressionskoeffizienten lässt sich die Stärke der einzelnen Risikofaktoren ermitteln. Anhand eines Beispiels wird das Grundprinzip der logistischen Regression sowie ihre Vorteile erläutert. Logistic regression is a model used for prediction of an event (e.g. heart attack: yes/no) using several predictor variables whereas the output is confined to values between 0 and 1. The regression coefficients describe the size of the contribution of the risk factor. In this short introduction an example is used to explain the basic principle and the advantages of the logistic regression.

DOI: https://doi.org/10.1024/1661-8157.97.14.779

Other titles: Identification of risk factors (the method of logistic regression)

Die Ermittlung von Risikofaktoren

Das Verfahren der logistischen Regression

Martin Sieber

3.4.2008

Zusammenfassung
Die logistische Regression ermöglicht die Bestimmung von Risikofaktoren für ein Zielkriterium, wobei das Zielkriterium zweistufig ist, z.B. krank/gesund. Aus den Regressionskoeffizienten lässt sich die Stärke der einzelnen Risikofaktoren ermitteln. Anhand eines Beispiels wird das Grundprinzip der logistischen Regression sowie ihre Vorteile erläutert.

Schlüsselwörter:
Logistische Regression – Risikofaktor – Odds – Odds Ratio

Summary
Logistic regression is a model used for prediction of an event (e.g. heart attack: yes/no) using several predictor variables whereas the output is confined to values between 0 and 1. The regression coefficients describe the size of the contribution of the risk factor. In this short introduction an example is used to explain the basic principle and the advantages of the logistic regression.

Key words:
Logistic regression – Risk Factor – Odds – Odds Ratio

Résumé
La régression logistique permet d'étudier la relation entre une variable réponse binaire (p.e. "success" ou "échec") et plusieurs variables explicatives. L'article donne une brève introduction à la régression logistique et illustre sa logique par un exemple.

Mots-clés:
Régression logistique – modèle de prédiction – risque – odds – odds ratios

Die logistische Regression ist ein Berechnungsverfahren, das in der Medizin häufig für die Bestimmung der Risikofaktoren einer Krankheit angewendet wird. Wie bei jeder Regressionsanalyse liegt ein wichtiger Vorteil dieses Verfahrens darin, dass die gegenseitige Beeinflussung der Risikofaktoren berücksichtigt wird. Der vorliegende Artikel soll eine Hilfe bieten, das Prinzip der logistischen Regression verständlich zu machen. An einem Beispiel wird aufgezeigt, wie mit Hilfe dieser Methode der Weg aus einer Sackgasse gefunden wurde.

Beispiel INTERHEART-Studie

Bei dieser Studie wurden 15 000 Personen mit akutem Myokardinfarkt und ebenso viele Kontrollpersonen aus 52 Ländern untersucht [1]. Mittels logistischer Regression wurden wichtige Risikofaktoren wie Rauchen, ungünstige Blutfette, hoher Blutdruck, Zuckerkrankheit, Fettleibigkeit und psychosozialer Stress ermittelt. Dabei konnte die Stärke der einzelnen Risikofaktoren unter Berücksichtigung von Alter und Geschlecht bestimmt werden. Mit Hilfe der logistischen Regression gewinnt man somit substantielle Aussagen zu den Risiken.

Bei der logistischen Regression sind drei Elemente wichtig: Die Risikofaktoren (Prädiktoren), das Zielkriterium (z.B. gesund/krank) und die Stärke oder Gewichtung der Risikofaktoren. In Grafik 1 sind oft verwendete Begriffe und ihre Relation zueinander dargestellt. In Tab. 1 werden deutsche und englische Fachbegriffe aufgeführt.

Etwa hier Grafik 1 einsetzen

Etwa hier Tabelle 1 einsetzen

Die logistische Regression ist mit der einfachen und der multiplen (mehrere Risikofaktoren) linearen Regression verwandt. Der Hauptunterschied liegt darin, dass bei der logistischen Regression das Zielkriterium zweistufig ist, z.B. krank/gesund, bei der einfachen oder multiplen Regression dagegen kontinuierlich wie z.B. Blutdruck, Körpergewicht. Allen gemeinsam ist das Ziel, die Stärke der einzelnen Risikofaktoren so zu ermitteln, dass damit das Zielkriterium (krank/gesund) am besten vorhergesagt werden kann. Die Bedeutung der einzelnen Risikofaktoren wird an der Höhe des sog. „Regressionskoeffizienen“ ersichtlich, die das Verfahren berechnet.

Zwei weitere Begriffe, die in Zusammenhang mit der logistischen Regression auftreten, sind die „Odds“ und die „Odds Ratio“. Das hat damit zu tun, dass wir beim Zielkriterium zwei Stufen haben (krank/gesund). Nehmen wir z.B. eine Gruppe von Rauchern, bei der eine bestimmte Anzahl einen Myokardinfarkt erlitt (krank), die übrigen jedoch nicht (gesund). Werden die beiden Zahlen als Bruch geschrieben, erhalten wir das Odds, ein für uns ungewohnter Begriff (siehe Kasten). Werte über 1.0 weisen auf ein Überwiegen von „krank“, Werte unter 1.0 auf ein Überwiegen von „gesund“, bei 1.0 ist das Verhältnis ausgeglichen.

Wenn wir nun die Odds für verschiedene Gruppen ermitteln, z.B. Raucher und Nichtraucher, dann können wir die Odds dieser Gruppen miteinander vergleichen und erfahren so, welche Teilgruppe das grössere Erkrankungsrisiko hat und ob das Rauchen ein Risikofaktor darstellt. Dieser Vergleich wird ebenfalls mit einer Zahl dargestellt, dem sog. „Odds Ratio“. Das Verhältnis der beiden Odds ist das Odds Ratio und ist in der logistischen Regression wichtig. Die Interpretation der Odds Ratio ist wie bei den Odds. Ziel ist, die Werte der Odds Ratios für die verschiedenen Risikofaktoren zu berechnen.

Etwa hier Kasten Odds einsetzen

Ein Beispiel: Ist Reserpin krebsfördernd?

Die Anwendung der logistischen Regression soll an einem Beispiel erläutert werden, das uns auch in die „Denkart“ der logistischen Regression und deren Vorzüge einführt. In den siebziger Jahren wurde diskutiert, ob das Medikament Reserpin bei Frauen Brustkrebs verursacht. Diese Kontroverse landete in einer Sackgasse, die mit Hilfe der logistischen Regression wieder verlassen werden konnte [2,3].

Die ersten Studienbefunde ohne Verwendung der logistischen Regression (Tab. 2) deuteten darauf hin, dass Reserpin eher einen krebsfördernden Einfluss hatte. Bei der Gruppe ‚mit Reserpin’ hatten 32 Frauen Brustkrebs, 57 jedoch nicht, was einem odds von 0.56 entspricht. Bei der Gruppe ‚ohne Reserpin’ hatten anteilmässig weniger Frauen Krebs (149 mit, 351 ohne Krebs), was ein odds von 0.42 ergibt. Um den Einfluss des Reserpins in Zahlen erfassen zu können, wird das Odds Ratio berechnet,

das einen Wert von 1.32 ergibt (0.56/0.42=1.32). Die Chance für Brustkrebs ist bei der Gruppe „mit Reserpin“ um 1.32 oder 32% grösser als ohne Reserpin.

Etwa hier Tab. 2 einsetzen

Kontroverse und Sackgasse

Mit diesem Resultat (Reserpin ist krebsfördernd) begann die Kontroverse. Von den Fachleuten wurde eingewendet, dass ältere Frauen häufiger Reserpin nehmen, um ihren Blutdruck zu reduzieren, der mit zunehmendem Alter ein Problem wird. Ferner sind es ebenfalls ältere Frauen, die an Brustkrebs leiden. Der erwähnte Befund könnte also das Resultat des Alters sein, einer Begleitvariable (confounder, intervenierende Variable), die in Tab. 2 nicht berücksichtigt wurde.

In Tab. 3 ist deshalb das Alter einbezogen worden. Wir sehen, dass bei den älteren Frauen viel mehr Brustkrebsfälle auftreten als bei den jüngeren. Betrachten wir zunächst lediglich die jüngere Gruppe (unter 50 J.), so erhalten wir ein Odds Ratio von 0.75 (2/14 dividiert durch 42/221). Die Einnahme von Reserpin vermindert das Brustkrebsrisiko um 25%, d.h. es liegt 0.25 unter dem Wert von 1.0.

Etwa hier Tab. 3 einsetzen

Bei der älteren Gruppe beträgt das Odds Ratio 0.85 und liegt ebenfalls noch unter 1.0. Innerhalb beider Altersklassen ergibt sich somit bei Reserpineinnahme eine Verringerung des Brustkrebsrisikos gegenüber 1.0 (kein Effekt) um 15% resp. 25%. Reserpin scheint prophylaktisch gegen Brustkrebs zu wirken. Dies ist ein deutlich anderes Ergebnis als die oben erwähnte Feststellung, wonach Reserpin ein Risikofaktor sei! Es bestanden nun zwei verschiedene Ergebnisse, man landete also in einer Sackgasse. Dabei wurde vermutet, dass das Alter und nicht das Reserpin der entscheidende Risikofaktor war, aber gesichert war das nicht. Um das ermitteln zu können bräuchte es eine Berechnungsart, bei der beide Risikofaktoren gleichzeitig berücksichtigt würden und ‚adjustierte’ Odds Ratios liefert.

Der Ausweg

Die logistische Regression bot die geeignete Hilfe, bei der beide Risikovariablen *simultan* einbezogen werden können. Bei unserem Beispiel möchten wir wissen, ob die Reserpineinnahme (X_1) ein signifikanter Risikofaktor für Brustkrebs ist, wenn das Alter (X_2) in die Berechnung einbezogen wird. Wir möchten also ein Odds Ratio für Reserpin erhalten, bei dem das Alter mitberücksichtigt wurde (adjustierte Odds Ratio).

Zweitens möchten wir wissen, welches Gesamtrisiko z.B. für ältere Frauen besteht, wenn sie kein Reserpin einnehmen, und welches Risiko vorliegt, wenn sie Reserpin einnehmen. Damit könnten wir Frauen, die uns diese Frage in der Praxis stellen, besser beraten.

Die Gleichung lautet somit:

$$Y = c + b_1 \cdot X_1 + b_2 \cdot X_2$$

Die Werte für X_1 und X_2 beziehen sich auf die Patientendaten. Wenn Reserpin eingenommen wurde, ist $X_1 = 1$, sonst 0. Wenn das Alter über 50 ist, ist $X_2 = 1$, sonst 0. Die b_1 und b_2 sind die Regressionskoeffizienten, welche die gesuchte Stärke des Risikofaktors angeben. Diese Regressionskoeffizienten, welche der Computer mit der sog. Maximum-Likelihood-Methode berechnet, sind allerdings nicht direkt interpretierbar und wir müssen auch hier eine Umwandlung vornehmen, damit wir zu den gesuchten Odds Ratios kommen, die wir dann interpretieren können. Dabei wird der natürliche Logarithmus verwendet. Von daher stammt die Bezeichnung „logistische“ Regression [4,5].

Die sogenannten logarithmierten Wettverhältnisse stellen in der logistischen Regressionsanalyse für $Y_i = 1$ eine lineare Funktion der Prädiktoren $x_i^{(j)}$ dar. D.h., es gilt:

$$\log (P(Yi = 1)/(1-P(Yi = 1))) = \beta_0 + \beta_1 x_i^{(1)} + \ldots \beta_m x_i^{(m)}$$

In unserem Beispiel lautet die Gleichung wie folgt:

$$\log (P(Y_i = 1)/(1-P(Y_i = 1))) = -1.67 - 0.18\, x_i^{(1)} + 1.47\, x_i^{(m)}$$

In der linearen Regressionsanalyse sagt der β-Wert aus, um wieviel die Zielvariable y zunimmt, wenn eine erklärende Variable x(j) um eine Einheit zunimmt. Bei der logistischen Regression übernimmt diese Rolle das sogennante Doppelverhältnis, dh. die sogenannte odds ratio. Erhöht man $x^{(j)}$ um eine Einheit, dann erhöht sich das logarithmierte Wettverhältnis log $(P(Y_i = 1)/(1-P(Y_i = 1)))$ um β, wobei alle anderen $x^{(k)}$ konstant gehalten werden. Mit anderen Worten, erhöht man in der logistischen Regression $x^{(j)}$ um eine Einheit, dann erhöht sich die Zielvariable Y um den Faktor e^{β}. In unserem Falle erhöht sich das Brustkrebsrisiko um 0.84, wenn Reserpin eingenommen wird. Das Brustkrebsrisiko erhöht sich um 4.37, wenn das Alter auf über 50 Jahre zu liegen kommt (Tab. 4, Spalte 3).

Etwa hier Tab. 4 einsetzen

Wie werden die Odds Ratio (e^b) interpretiert? Die Interpretation ist so einfach wie bei den Odds: Werte über 1 bedeuten ein Risiko, Werte unter 1 entsprechen einer Risikoverminderung, 1.0 entspricht der neutralen Position. Der Wert von Odds Ratio OR=0.84 liegt somit unter 1.0 und besagt, dass Reserpin in der Tendenz risikovermindernd wirkt. Ist dieser Effekt signifikant? In den meisten Publikationen wird dazu entweder die Signifikanz oder das Konfidenzintervall (Vertrauensintervall) angegeben. Beim Konfidenzintervall geben uns die beiden Grenzwerte darüber Bescheid, ob das Merkmal signifikant ist.

Signifikanz

Wie wir in Tab. 4 sehen ergibt sich für die Reserpin-Einnahme ein altersadjustiertes Brustkrebs-Risiko von OR=0.84. Reserpin wirkt, unter Berücksichtigung des Alters, prophylaktisch, der Effekt ist jedoch nicht signifikant. Dies geht aus der Spalte rechts aussen hervor, die das sog. Konfidenzintervall des Odds Ratio angibt. Das 95%-Konfidenzintervall hat zwei Grenzwerte und gibt an, wie stark die Werte schwanken, wenn theoretisch 100 gleiche Studien durchgeführt würden. Bei 95 der 100 Studien würde das Odds Ratio zwischen dem unteren Grenzwert (0.51) und dem oberen Grenzwert (1.38) liegen. Ein Wert ausserhalb dieser Grenzwerte kommt sehr selten vor (in nur 5 von 100 Studien), so dass ein solcher Wert als signifikant bezeichnet würde. Beim vorliegenden Beispiel sehen wir, dass der Wertebereich für das Konfidenzintervall

von 0.51 bis 1.38 geht und somit 1.0 (kein Effekt) einschliesst. Es ist also ziemlich wahrscheinlich, dass „kein Effekt des Reserpins“ vorliegt. Weil nun das Ergebnis „kein Effekt“ nicht ausgeschlossen werden kann, wird das Ergebnis als „nicht signifikant“ bezeichnet. Reserpin hat keinen statistisch signifikanten Einfluss.

Etwa hier Kasten Vertrauensintervall einsetzen

Bezüglich des Alters sieht die Situation anders aus. Für erhöhtes Alter ergibt sich ein Reserpin-adjustiertes Brustkrebs-Risiko von OR=4.37 mit den Konfidenzgrenzen von 2.92 und 6.55. Dies besagt, dass das Brustkrebsrisiko bei älteren Frauen um den Faktor 4.37 gegenüber jüngeren Frauen erhöht ist. Die untere Grenze des 95%-Konfidenzintervalles ist erheblich grösser als 1, nämlich 2.92. Dieser Wert ist also deutlich höher als der neutrale Wert 1.0 (kein Effekt). Es kann deshalb mit hoher Wahrscheinlichkeit ausgeschlossen werden, dass kein Effekt vorliegt, mit anderen Worten: Der Effekt ist signifikant. Manchmal wird in den Studien das Konfidenzintervall nicht erwähnt, jedoch die Signifikanz der Risikovariabeln aufgeführt, was zum gleichen Ergebnis führt. Die statistische Signifikanz der einzelnen Risikofaktoren kann mit dem sog. WALD-Test geprüft werden.

Wie erwähnt können mit Hilfe der logistischen Regression auch die Risiken einzelner Patientengruppen berechnet werden. Dazu werden in der Gleichung mit dem Risiko-Summenscore X die entsprechenden Ladungen von Tab. 4 eingesetzt. Für die Gruppe „Alter über 50, kein Reserpin“ erhalten wir ein logarithmiertes Wettverhältnis von -0.20. Die Umwandlung in einen %-Wert (hier nicht ausgeführt) ergibt ein Risiko von 45.0%. Bei der Gruppe „Alter über 50, mit Reserpin“ erhalten wir ein logarithmiertes Wettverhältnis von -0.38, was nach der Umwandlung einem Risiko von 40.6% entspricht. Mit Reserpin kann also das Brustkrebsrisiko um 4.4% gesenkt werden. Auch bei den jüngeren Frauen kann das Brustkrebsrisiko mit Reserpin gesenkt werden, allerdings nur um 2.2%. Mit einem Anpassungstest wird geprüft, ob eine bestimmte Verteilung eine gegebene Stichprobe genügend gut beschreibt. In unserem Falle wollen wir überprüfen, ob ein Modell mit nur einem Parameter (z.B. Reserpin) das richtige ist (H_0-Hypothese) oder ob das maximale Modell mit 2 Parametern (Reserpin und Alter) das richtige ist (H_1). Dabei wird die Null-Devianz mit der Residuendevianz verglichen.

Key messages

Die logistische Regression ist ein Berechnungsverfahren, das für die Bestimmung der Risikofaktoren einer Erkrankung angewendet wird.

Mit ihr kann die Stärke und die Signifikanz dieser Risikofaktoren bestimmt werden.

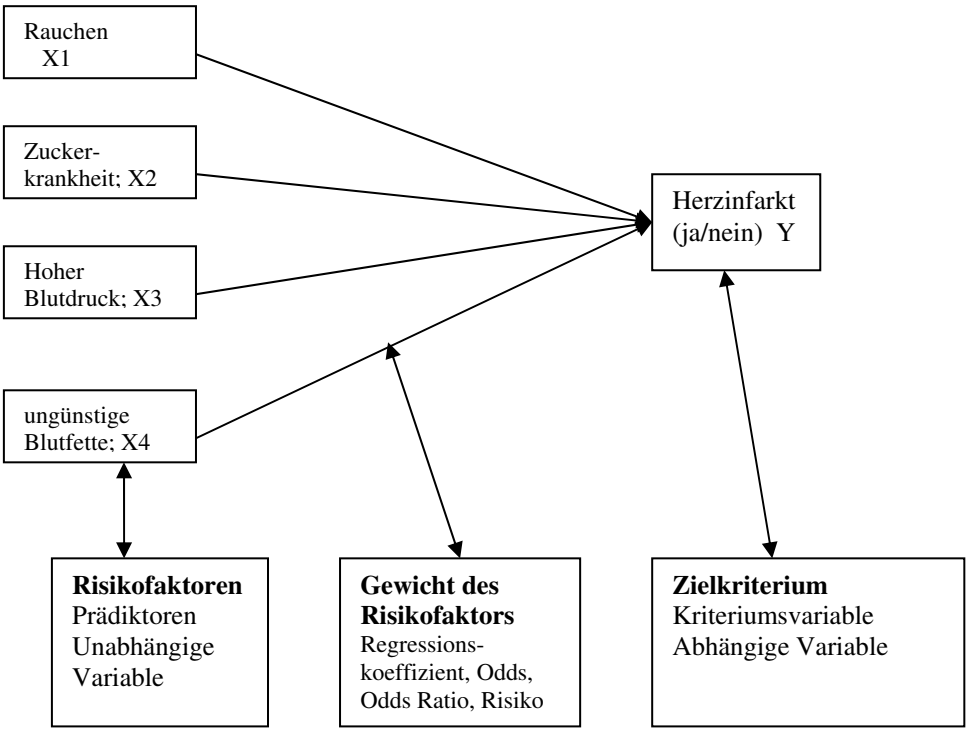
Ein wichtiger Vorteil liegt wie bei anderen Regressionsverfahren darin, dass die gegenseitige Beeinflussung der Risikofaktoren berücksichtigt wird.

Literatur

1. Yusuf S, Hawkwn S, Ounpuu S et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case control study. Lancet 2004; 364: 912-4
2. Kewitz H, Jesdinsky HJ, Schröter PM, Lindtner E. Reserpine and Breast Cancer in Women in Germany. Europ J clin Pharmacol. 1977; 11: 79-83
3. Kewitz H, Härter G, Feldmann U, Kreutz G, Nitz M, Unger E. Observational Cohort Study in General Practice: Differences and Equivalences Among Analgesics for Treatment of Colic Pain. In: Kewitz H. (ed): Epidemiological Concepts in Clinical Pharmacology. New York: Springer, 1987, pp 73-86
4. Bender R, Ziegler A, Lange St. Logistische Regression. Dtsch Med Wochenschr 2002; 127: T11-T13
5. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley, 1989

Korrespondenzadresse:
Prof. Dr. phil. Martin Sieber
Zentrum für Zahn-, Mund- und Kieferheilkunde der Universität Zürich, Plattenstrasse 11, CH 8032 Zürich.

martin.sieber@zzmk.uzh.ch.



Grafik 1: Beispiel für das Zusammenspiel des Zielkriteriums mit den Risikofaktoren

Tabelle 1: Wichtige Begriffe in Deutsch und Englisch

Deutsch	Englisch
Logistische Regression	Logistic regression
Einfache (multiple) lineare Regression	Simple (multiple) linear regression
Erklärende Variable, Risikofaktor	Explanatory factor, risk factor
Zielvariable	Response variable, dependent variable
Regressionskoeffizient	Regression coefficient
Chance	Odds
Odds-Verhältnis, relative Chance	Odds Ratio
Vertrauens- oder Konfidenzintervall	confidence interval, C.I.
Verzerrung	Bias
Adjustiert	Adjusted
Wechselwirkung	Interaction
Modellgüte	Goodness-of-fit

Kasten Odds

Odds wird als Chance oder (umgangssprachlich) als Risiko übersetzt. Mathematisch berechnen sich Odds als Quotient aus der Wahrscheinlichkeit, dass ein Ereignis eintritt und der Wahrscheinlichkeit, dass es nicht eintritt: $O_i = P_i / (1 – P_i)$. Ein Beispiel: Zieht man bei einem Kartenspiel eine Karte, dann beträgt die Odds für ein "Herz" 0.25 / 0.75 = 1/3. Oder ein Münzenbeispiel: Die Odds für "Kopf" auf der Münze ist 0.5 / 0.5 = 1. Odds können Werte zwischen Null (das Ereignis tritt nie ein) und Unendlich (tritt mit Sicherheit ein) annehmen; 1 ist die neutrale Position. Das Odds Ratio ist das Verhältnis von zwei Odds: $OR = O_1 / O_2$. Es ist verwandt mit dem Relativen Risiko, aber nicht genau identisch.

Tab. 2: Risiko und Odds bei Studie „Reserpineinnahme und Brustkrebs“

Reserpin-Einnahme	Brustkrebs Ja	Brustkrebs Nein	Total	Odds
Ja	32	57	89	0.56
Nein	149	351	500	0.42
Gesamt	181	408	589	0.44

Tab. 3: Aufteilung der Patientinnen in zwei Altersgruppen (bis 50; über 50 J.)

Alter	bis 50 Jahre		über 50 Jahre	
	Brustkrebs Ja	Brustkrebs Nein	Brustkrebs Ja	Brustkrebs nein
Reserpin Ja	2	14	30	43
Reserpin Nein	42	221	107	130
Gesamt	44	235	137	173

Tab. 4: Ergebnis der logistischen Regression

Variable	Wert der Ladung	Odds Ratio (e^b)	95%-Konfidenzintervall
B_1 (Reserpin)	- 0.18	OR= $e^{-0.18}$ = 0.84	0.51; 1.38
B_2 (Alter)	+1.47	OR= $e^{1.47}$ = 4.37	2.92; 6.55
C (Konstante)	- 1.67	OR= $e^{-1.67}$ = 0.19	0.14; 0.26

Kasten: Vertrauensintervall, Konfidenzintervall (K.I., C.I.)

Das Konfidenzintervall ist der Unsicherheitsbereich für die Schätzung eines bestimmten, nicht bekannten Parameters. Das 95%-Konfidenzintervall kennzeichnet denjenigen Bereich eines Merkmals, in dem sich 95% aller möglichen Werte befinden. Ein Odds Ratio ist signifikant, wenn der Wert 1.0 ausserhalb des Konfidenzintervalls liegt.