CHEST - Ein SAS-Macro für konventionelle und Change-in-Estimate basierte Variablenselektion zur Modellierung epidemiologischer Daten

U. Siebert[(1)], N. Mühlberger[(2)], A. Wulff[(2)]

(1) Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie, Ludwig-Maximilians-Universität München

(2) GSF-Forschungszentrum für Umwelt und Gesundheit, Institut für Medizinische Informatik und Systemforschung, Neuherberg

Abstract

Hintergrund: Es gibt zwei Ziele der multivariaten Modellierung: die effektive Prädiktion der Zielgröße und die valide Quantifizierung des Effektes einer bestimmten Exposition auf die Zielgröße. Letzteres ist die Zielsetzung in epidemiologischen Risikofaktorstudien. Für die Effektquantifizierung von Risikofaktoren wurde von verschiedenen Epidemiologen die Verwendung des Change-in-Estimate Kriteriums anstelle der statistischen Signifikanz als maßgebliches Kriterium für die Selektion von Confoundern vorgeschlagen. Der von den meisten Softwarepaketen angebotene konventionelle Stepwise-Regressionsalgorithmus basiert ausschließlich auf der Signifikanz der Assoziation zwischen den Covariablen und der Zielgröße und berücksichtigt damit nicht das zweite klassische Confounderkriterium, die Assoziation der Covariable mit der Exposition. Das Change-in-Estimate Kriterium ist in den erhältlichen einschlägigen Statistiksoftwarepaketen nicht als Option für Variablenselektionsverfahren implementiert.

Das CHEST-Makro: Das von uns entwickelte SAS-Makro CHEST für die multiple logistische Regression ermöglicht die Wahl zwischen drei verschiedenen Methoden der Variablenselektion. Die erste Methode ist die konventionelle, auf der Signifikanz der Kandidatenvariablen basierende Strategie. Die zweite Methode basiert auf der relativen Änderung der Odds Ratio beim Vergleich der Modelle mit und ohne jeweils untersuchte Kandidatenvariable. Die dritte Methode basiert auf dem Collapsibility Test, der einen Signifikanztest für die relative Änderung der Odds Ratio darstellt. Jede Methode kann im Forward- oder Backward-Verfahren bei beliebig zu definierenden Schrankenwerten ausgeführt werden. Bei jedem Schritt der Variablenselektion wird ein tabellarischer Ausgabebericht generiert, der unter Verwendung der jeweils maximal zur Verfügung stehenden Daten die Ergebnisse aller drei Methoden der Variablenselektion darstellt.

Die praktische Anwendbarkeit von CHEST wird anhand des Public-Use-Datensatzes der Framingham-Studie dargestellt. Vorzüge und Grenzen des Change-in-Estimate als Kriterium für die Selektion von Confoundern werden aufgezeigt.

Schlußfolgerungen: CHEST stellt ein systematisches und effizientes Instrument für die Selektion multipler Confounder dar, insbesondere für Situationen, in denen nur limitiertes Vorwissen über die Confoundingstruktur vorhanden ist. Zur systematischen Untersuchung der Effektivität der einzelnen Selektionsstrategien sind Simulationsstudien mit multiplen Covariablen erforderlich. Die Implementierung des CHEST-Makro ermöglicht eine effiziente Durchführung solcher Simulationsstudien.

Hintergrund

Multivariate Modelle werden in der Epidemiologie für zwei Ziele eingesetzt: (i) für die effektive Prädiktion der Zielgröße oder (ii) die valide Quantifizierung des Effektes einer bestimmten Exposition auf die Zielgröße [1,2]. Das erste Ziel wird in Diagnose- und Prognosestudien, das zweite in epidemiologischen Risikofaktorstudien verfolgt. Dieser Beitrag beschränkt sich ausschließlich auf das Ziel der objektiven und validen Effektquantifizierung und stellt kurz die Probleme dar, die mit den verschiedenen Selektionsstrategien für Confounder verbunden sind. Der von den meisten Softwarepaketen angebotene konventionelle Stepwise-Regressionsalgorithmus basiert ausschließlich auf der Signifikanz der Assoziation zwischen den Covariablen und der Zielgröße und berücksichtigt damit nicht das zweite klassische Con-

founderkriterium, die Assoziation des Confounders mit der Exposition. Aus diesem Grund wurde von verschiedenen Epidemiologen die Verwendung des Change-in-Estimate Kriteriums anstelle der statistischen Signifikanz als maßgebliches Kriterium für die Selektion von Confoundern vorgeschlagen [1-4]. Leider sind in den erhältlichen einschlägigen Statistiksoftwarepaketen solche Algorithmen nicht implementiert. Eine systematische manuelle Durchführung dieser Algorithmen ist zwar möglich, jedoch zeitaufwendig und fehleranfällig.

Das CHEST-Makro

Das von uns entwickelte SAS-Makro CHEST für die multiple logistische Regressionsanalyse ermöglicht die Wahl zwischen drei verschiedenen Methoden der Variablenselektion. Die Optionen des Makros sind in Abb. 1 dargestellt.

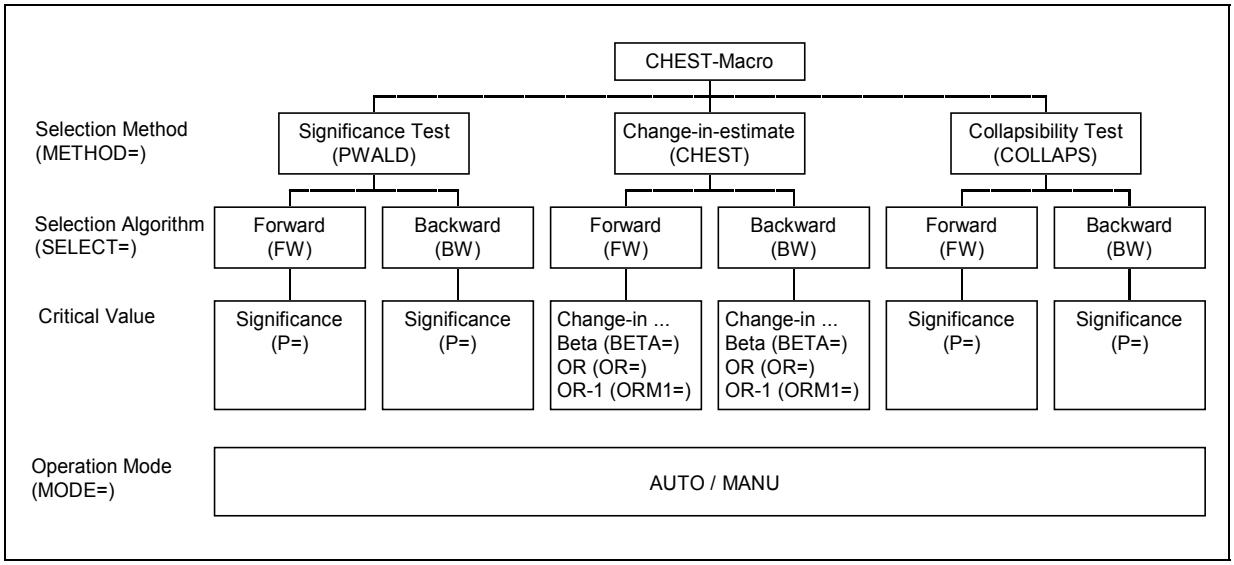

Abbildung 1: Überblick über die Optionen des CHEST-Makro: Selektionsverfahren (Selection Method), Stepwise-Algorithmen (Stepwise Algorithm), kritische Schranken (Critical Value) und Durchführungsmodi (Operation Mode).

Die erste Methode (PVALUE) ist die konventionelle, auf der Signifikanz (P) der Kandidatenvariablen basierende Strategie. Die zweite Methode (CHEST) basiert auf der relativen Änderung der Odds Ratio beim Vergleich der Modelle mit und ohne jeweils untersuchte Kandidatenvariable [4-6]. Aus Gründen, auf die in der Diskussion eingegangen wird, wurden zwei weitere Varianten der CHEST-Methode implementiert: die Verwendung der relativen Änderung des Regressionskoeffizienten (BETA) und die relative Änderung der Odds-Ratio-minus-eins Skala (ORM1). Alle Change-in-Estimate Kriterien basieren auf folgender Formel:

$$\text{Relative Änderung} = \left\lfloor \exp\left(\left| \log(\hat{\theta}_{adj}) - \log(\hat{\theta}_{crude}) \right| \right) - 1 \right\rfloor \times 100\%, \quad \text{mit } \hat{\theta} = \text{beta, OR oder OR-1.}$$

Die dritte Methode (COLLAPS) basiert auf dem Collapsibility-Test von Greenland und Mikkey [5] der einen Test auf Signifikanz (P) der relativen Änderung der Odds Ratio darstellt und die Collapsibility Testgröße nach Wald verwendet:

$$W = \frac{\hat{\beta}_{adj} - \hat{\beta}_{crude}}{\sqrt{Var(\hat{\beta}_{adj}) - Var(\hat{\beta}_{crude})}},$$

wobei W^2 Chi^2-verteilt ist und W=0 falls $Var(\hat{\beta}_{crude}) \geq Var(\hat{\beta}_{cadj})$.

Jede Methode kann im Forward-Verfahren (FW) oder im Backward-Verfahren (BW) bei beliebig zu definierenden Schrankenwerten für die Signifikanz der Kandidatenvariablen (P) bzw. beliebig zu definierenden Schrankenwerten für den relativen Change-in-Estimate (BETA, OR, ORM1) ausgeführt werden.
Der Programmodus kann von automatisch (AUTO) auf manuell (MANU) umgeschaltet werden. Im automatischen Modus wird schrittweise die Variablenselektion solange durchgeführt, bis das Endmodell erreicht ist. Der manuelle Modus kann eingesetzt werden, wenn die Struktur der fehlenden Werte des Datensatzes dies gebietet. In diesen Modus stoppt das CHEST-Makro nach jedem Modellierungsschritt und erlaubt damit dem Analysierenden, in den Modellbildungsprozeß einzugreifen. Im Gegensatz zu automatischen Prozeduren in kommerziellen Softwarepaketen ist das CHEST-Makro bei den schrittweise ausgeführten Variablenselektionsverfahren nicht auf eine Complete-Case-Analyse beschränkt, sondern ermöglicht selbst im automatischen Modus die maximale Nutzung der Daten bei jedem Modellierungsschritt, d.h. bei jedem Vergleich von Modellen. Weitere Argumente des Makros sind die Spezifikation der SAS-Datendatei (FILE), der Zielgröße (OUT), der Exposition (EXPO), der fest (forced-in) ins Modell aufzunehmenden Covariablen (INCLUDE) und der Kandidaten-Covariablen für den Selektionsprozeß (COVAR).
Bei beiden Durchführungsmodi wird für jeden Modellierungsschritt ein tabellarischer Ausgabebericht generiert, der die Ergebnisse aller drei Methoden der Variablenselektion darstellt. Die Variablen werden nach der Größe des ausgewählten Selektionskriteriums sortiert aufgelistet. Tab. 1 zeigt die Anwendung des CHEST-Makro auf den Framingham Public-Use-Datensatz. Analysiert wird der Effekt der Hypertonie (HPT) auf die kumulative Inzidenz von koronaren Herzerkrankungen (CHDNEW). Als potentielle Confounder wurden Geschlecht (SEX), Alter (AGE), Zigarettenkonsum (CIG) und Cholesterinwert (CHOL) berücksichtigt.

Tabelle 1: Anwendung des CHEST-Makro auf den Framingham Public-Use-Datensatz: Output für eine auf dem Change-in-OR Kriterium basierende Forward-Selektion.

```
01: C H E S T  1.0
02: FILE    = FRAM      N = 1363
03: OUTCOME = CHDNEW    EXPOSURE  = HPT        INCLUDE =      COVARIABLES = AGE SEX CHOL CIG
04: METHOD  = CHEST     SELECTION = FORWARD    OR   = 10%     MODE = AUTOMATIC
05:-------------------------------------------------------------------------------------------
06: STEP 1:   MODEL: CHDNEW = HPT + X
07:-------------------------------------------------------------------------------------------
08: OBS  Covariable  p-value   N       Beta     Beta     ΔBeta    %Beta    %OR   %(OR-1)  p-value
09:      (X)         (Wald)            crude    adj.                                      (Collaps)
10:-------------------------------------------------------------------------------------------
11:   1   HPT         0.0001   1363    0.674     .        .        .       .      .       .
12:   2   CHOL        0.0220   1363    0.674    0.652    0.022     3.3     2.2     4.6    0.0780
13:   3   CIG         0.0014   1362    0.677    0.712   -0.035     5.2     3.6     7.2    0.0567
14:   4   AGE         0.0001   1363    0.674    0.636    0.038     6.0     3.9     8.2    0.0313
15:   5   SEX         0.0001   1363    0.674    0.745   -0.071    10.5     7.4    15.0    0.0100
```

Kommentar zu Tab.1:
In Zeile 02-04 sind die Makro-Argumente und die Fallzahl dargestellt. Zeile 06 zeigt die zu vergleichenden Modelle bei dem aktuellen Modellierungsschritt, Zeile 11 den rohen Expositi-

onseffekt und Zeile 12-15 die rohen und adjustierten Expositionseffekte und die zugehörigen Maße für den jeweiligen Modellvergleich auf Basis der maximal verwendbaren Daten.

Diskussion

Es besteht Konsens über die Tatsache, daß eine adäquate Confounder-Selektion auf Substanzwissen der jeweiligen Disziplin basieren sollte und vor Beginn der Datenanalyse eine gründliche Sichtung der Literatur bezüglich der gegebenen Fragestellung erforderlich ist. Bekannte Confounder sollten ins Modell aufgenommen werden und Intermediärvariablen in der Kausalkette zwischen Exposition und Zielgröße sollten bei der Analyse von Kausalfaktoren ausgeschlossen werden [1,2]. Allerdings ist bei epidemiologischen Fragestellungen aufgrund komplexer biologischer und medizinischer Mechanismen häufig nur limitiertes Vorwissen über Confounding in der untersuchten Population gegeben und eine datengestützte Confounderselektion ist unvermeidlich.
Die Limitationen der signifikanzbasierten Confounderselektion wurden in der Literatur der letzten 20 Jahre mehrfach detailliert herausgestellt und diskutiert [1-4]. Limitationen des Change-in-Estimate Selektionsverfahrens für Confounder sind auf fehlende Guidelines für nicht-dichotome Expositionsvariablen und für Effektmodifikation zurückzuführen. Während für dichotome Variablen ein Change-in-OR von 10% als Selektionskriterium vorgeschlagen wurde [5], ist dieses Vorgehen für kontinuierliche Expositionsvariablen nicht angemessen, da die relative Änderung der OR stark von der gewählten Skaleneinheit der Expositionsvariablen abhängt. Für diesen Fall schlagen wir als Selektionskriterium die relative Änderung des Exzeß-Risikos (OR-1) oder die relative Änderung des logistischen Regressionskoeffizienten vor. Beide Varianten wurden im CHEST-Makro realisiert (ORM1, BETA). Eine weitere Limitation des Change-in-Estimate Ansatzes besteht darin, daß kategoriale Expositionsvariablen oder die Anwesenheit von Effektmodifikation zu einem Vergleich multipler Effektschätzer führen [1]. Für diese Situationen sind zusammenfassende Kriterien zu definieren, z.B. die Berücksichtigung der maximalen oder des mittleren (gepoolten) Änderung des Effektes. Ferner wurde darauf hingewiesen, daß die Identifikation eines adäquaten Subsets von Confoundern aus zwei Gründen eine Backward-Eliminationsstrategie erfordert: erstens sollte der Wert für Change-in-Estimate auf ein volladjustiertes Modell bezogen werden und zweitens kann aufgrund der Abwesenheit von Marginal-Confounding nicht notwendigerweise auf die Abwesenheit von Joint-Confounding geschlossen werden [1]. Die Implikation hierbei ist die Annahme, daß die volle Adjustierung zur geringsten Verzerrung des Effektschätzers führt. Es wurde allerdings nachgewiesen, daß eine Adjustierung seinerseits zu einer Verzerrung des Effektschätzers führen kann, wenn Informationsfehler vorliegt oder die Covariablen nicht angemessen kategorisiert werden [7]. Eine Einschränkung aller Stepwise-Algorithmen liegt in der Tatsache, daß Modellvergleiche nur jeweils zwischen einem und dem nächsten Schritt angestellt werden und diese Algorithmen somit aggregiertes Confounding über mehrere Schritte hinweg nicht berücksichtigen. Für die Change-in-Estimate basierte Backward-Variablenelimination empfehlen wir, das Startmodell als 'Ankermodell' zu verwenden. Dies bedeutet, den Variablenabbau genau dann zu stoppen, wenn das über alle Abbauschritte kumulierte Change-in-Estimate bezüglich des vollen Modelles den gewählten kritischen Wert erreicht.
Zusammenfassend liefert das CHEST-Makro einen wichtigen Einblick in den Modellierungsprozeß und ermöglicht die Beobachtung sowohl der signifikanzbasierten als auch der Change-in-Estimate basierten Kriterien bei der Modellierung epidemiologischer Daten. Verschiedene Selektionsstrategien können als Sensitivitätsanalysen eingesetzt werden, um eine erhöhte Aussagekraft bezüglich des Effektes der untersuchten Exposition auf die Zielgröße zu erhalten.

Schlußfolgerungen

CHEST stellt ein systematisches und effizientes Instrument für die Selektion multipler Confounder dar, insbesondere für Situationen, in denen nur limitiertes Vorwissen über die Confoundingstruktur vorhanden ist. Guidelines für die Analyse nicht-dichotomer Expositionsmerkmale und für den Umgang mit Effektmodifikation sind bislang nicht hinreichend definiert.
Zur systematischen Untersuchung der Performanz der einzelnen Selektionsstrategien sind Simulationsstudien mit multiplen Covariablen erforderlich (Details sind dem von Mühlberger et al. eingereichten Beitrag in diesem Band zu entnehmen). Die Implementierung des CHEST-Makro ermöglicht eine effiziente Durchführung solcher Simulationsstudien.

Referenzen

[1] Kleinbaum D.G., Kupper L.L., Morgenstern H.: Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Lifetime Learning Publications 1982.

[2] Rothman K.J.: Modern Epidemiology. Boston/Toronto: Little Brown and Company 1986.

[3] Miettinen O.S.: Stratification by a multivariate confounder score. Am J Epidemiol 104, 1976, 609-620.

[4] Greenland S.: Modeling and variable selection in epidemiologic analysis. Am J Public Health 79, 1989, 340-349.

[5] Mickey RM, Greenland S.: The impact of confounder selection criteria on effect estimation. Am J Epidemiol 129, 1989, 125-137.

[6] Greenland S., Rothman K.J.: Introduction to stratified analysis. In: Greenland S., Rothman K.J. (eds.): Modern Epidemiology. Philadelphia, PA: Lippincott-Raven 1998.

[7] Brenner H.: A potential pitfall in control of covariates in epidemiologic studies. Epidemiology 9, 1997, 68-71.

Korrespondenzadressen für Tester des CHEST-Makro

Dr. Nikolai Mühlberger, M.P.H postgrad.
GSF-Forschungszentrum für Umwelt und Gesundheit
Institut für Medizinische Informatik und Systemforschung
Ingolstädter Landstraße 1
D-85764 Neuherberg
Tel.: 089-3187-4590
Email: muehlberger@gsf.de

Uwe Siebert, M.P.H postgrad.
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie
Ludwig-Maximilians-Universität München
Marchioninistr. 15
D-81377 München
Tel.: 089-7095-4482
Fax: 089-701000
Email: sieb@lrz-muenchen.de