Validierung von Regressionsmodellen: Notwendigkeit und Beschreibung der wichtigsten Methoden

The Necessity of Validating Regression Models – Characterization of the Most Important Methods

Autor R. Muche

Institut Institut für Biometrie, Universität Ulm

Schlüsselwörter
- Validierung
- Optimismus
- Kreuzvalidierung
- Bootstrap
- Shrinkage

Key words
- model validation
- optimism
- cross-validation
- bootstrap
- shrinkage

Koordinatoren der Reihe „Methoden in der Rehabilitationsforschung“:
Prof. Dr. Dr. Hermann Faller, Würzburg; Prof. Dr. Thomas Kohlmann, Greifswald; Dr. Christian Zwingmann, Siegburg
Interessenten, die einen Beitrag zur Reihe beisteuern möchten, werden gebeten, vorab Kontakt aufzunehmen, E-Mail: christian.zwingmann@web.de

Bibliografie
DOI 10.1055/s-2008-1077068
Rehabilitation 2008; 47: 243–250
© Georg Thieme Verlag KG Stuttgart · New York
ISSN 0034-3536

Korrespondenzadresse
PD Dr. Rainer Muche
Universität Ulm
Institut für Biometrie
Schwabstr. 13
89075 Ulm
rainer.muche@uni-ulm.de

Zusammenfassung

In der rehabilitationswissenschaftlichen Literatur werden Jahr für Jahr Studienergebnisse beschrieben, die auf Ergebnissen aus Regressionsanalysen beruhen. Der Einsatz dieser Regressionsmodelle erfolgt aber selten für die Diagnose oder Prognose des Patienten-Outcomes. Dies liegt in vielen Fällen daran, dass die sinnvolle Anwendung im rehabilitationsmedizinischen Alltag oft nicht untersucht wurde. Eine gute Modellierung des zu untersuchenden Zusammenhanges kann nicht immer auf weitere Patienten übertragen werden, da sich diese hinsichtlich wichtiger Variablen unterscheiden können. Deshalb ist es notwendig, Untersuchungen in Bezug auf diese Übertragbarkeit durchzuführen. Dies wird als Modellvalidierung bezeichnet. Wesentlicher Aspekt der Modellvalidierung ist die Untersuchung des Vorhersagefehlers, der durch eine zu optimistische Schätzung der Regressionseigenschaften erzeugt wird. Dieser Effekt entsteht durch die zweifache Nutzung des Datensatzes, einmal zur Modellentwicklung und dann zur Beurteilung der Eigenschaften. Dieser systematische Fehler (Bias) wird im Beitrag genauer beschrieben. Darüber hinaus wird eine Übersicht über mögliche Validierungsmethoden wie Kreuz- und Bootstrap-Validierung gegeben. Ein Vergleich der Verfahren und ein Beispiel schließen die Beschreibung ab.

Abstract

In rehabilitation research regression models are often used in analysing the effects of several independent factors on important outcomes in rehabilitation. But the results of such models are rarely used in rehabilitation practice for diagnosis or prognosis of patient outcomes. The main reason for this is the fact that the applicability of such models for new patient data is often unknown. A good fit of the model with respect to the data used in modeling does not guarantee a well-fitting model in the future. Thus, it is necessary to examine the performance of the model for new patient data. This examination is called model validation. The main aspect in model validation is the investigation of the prediction error caused by a too optimistic estimation of the model parameters. This prediction error is due to the twofold use of the data set: for estimating the regression coefficients *and* for examining the model fit. In this paper this error is discussed and the main methods for regression model validation are presented. Finally an example illustrates the effects of model validation for prognostic purposes in a logistic regression model estimating the risk of early retirement after in-patient rehabilitation.

Einleitung

Eine der Erfahrungen in den Methodenzentren der Rehabilitationswissenschaftlichen Forschungsverbünde in den letzten Jahren [1] bei der Betreuung der Analyse rehabilitationswissenschaftlicher Studien ist, dass die Regressionsanalyse bei der Auswertung oft zum Einsatz kommt. Häufig wird zur Klärung von Zusammenhängen zwischen Messgrößen diese Analysemethode benutzt, da so Störgrößen und weitere Einflussgrößen berücksichtigt werden können. Dabei wird in der Regressionsanalyse untersucht, inwiefern Ausprägungen einer abhängigen Zielgröße sich erklären lassen durch die Ausprägungen einer oder mehrerer unabhängiger Variablen. Somit können Modelle aufgestellt werden, um interessierende Zielgrößen anhand bekannter Mess-

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

werte schätzen zu können. In der Literatur werden zwar die Ergebnisse von Regressionsmodellen beschrieben, aber deren Einsatz zur Schätzung der Zielgrößen in der realen Anwendung erfolgt eher selten [2]. Man verlässt sich bei der Diagnose, Therapieempfehlung und/oder Prognose eines Patienten lieber auf andere Informationen als auf einen so berechneten Modellwert. Dies liegt in vielen Fällen daran, dass die sinnvolle Anwendung im rehabilitationsmedizinischen Alltag oft nicht untersucht wurde. Eine gute Modellierung des zu untersuchenden Zusammenhanges kann nämlich nicht immer auf weitere Patienten übertragen werden. Deshalb ist es notwendig, Untersuchungen in Bezug auf diese Übertragbarkeit anzustellen. Die Überprüfung des Modellverhaltens in neuen Datensituationen wird als Modellvalidierung bezeichnet.

Der Begriff Validierung stammt ursprünglich aus dem Bereich der Qualitätssicherung. Laut Definition gemäß FDA (Food and Drug Administration, Zulassungsbehörde für Arzneimittel in den USA) bezeichnet der Begriff Validierung die „Tätigkeit, die sich mit den Teilprozessen des Testens und der Freigabeentscheidung befasst" ([3], S. 17). Danach ist für die Modellvalidierung zu zeigen, dass das entwickelte Regressionsmodell die notwendigen Eigenschaften besitzt (Test) und dass diese Eigenschaften auch für den Einsatz in der Praxis gut genug sind (Freigabe).

Altman und Royston beschreiben in dem mit „What do we mean by validating a prognostic model?" überschriebenen Artikel [4], was bei einer Validierung des Modells untersucht werden sollte: Modellstruktur, Design der Studie und „Transportabilität" des Modells. Außerdem legen sie folgende Stufen der Validierung fest: intern und extern (unabhängige Evaluation). Folgende Fragen werden von ihnen gestellt:

- Sind dieselben Variablen nach Validierung wichtig?
- Ist die funktionale Form des Modells korrekt?
- Sind die geschätzten Regressionskoeffizienten vergleichbar?
- Wie gut ist die Modell-Güte für die neuen Daten?

Harrell [5] beschreibt den Sinn einer Modellvalidierung folgendermaßen: „Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects or subjects not used to develop our model" (S. 90).

Wesentlicher Aspekt der Modellvalidierung ist also die Untersuchung des Vorhersagefehlers, der durch eine zu optimistische Schätzung der Eigenschaften eines Modells in Bezug auf die Prognose erzeugt wird. Dieser Effekt entsteht, wenn der Datensatz zweifach genutzt wird, einmal zur Modellentwicklung und dann zur Beurteilung der Eigenschaften. Dieser systematische Fehler (Bias), von Efron und Tibshirani als Optimismus bezeichnet [6], wird im anschließenden Kapitel genauer beschrieben. Danach wird eine Übersicht über mögliche Validierungsmethoden gegeben. Ein Vergleich der Verfahren und ein Beispiel schließen die Beschreibung ab.

Betont werden kann in diesem Zusammenhang aber nicht oft genug, dass es aufgrund dieses Optimismus bei der Schätzung von klinik- oder rehabilitationsrelevanten Daten auf der Basis eines Regressionsmodells dringend notwendig ist, dieses Modell vor dem Einsatz an neuen Daten in der Praxis zu validieren. Die Therapie- und sonstigen Entscheidungen, die auf Basis eines Modells getroffen werden, können für Patienten weitreichende Folgen haben und sollten auf möglichst sicherer Grundlage erfolgen.

Modellgüte

Im folgenden Abschnitt „Definition der Modellgüte" wird zuerst der Begriff Modellgüte eingeführt, der eine zentrale Bedeutung für die Validierung von Regressionsmodellen hat. Anschließend wird im Abschnitt „Bias der Modellgüte" das Problem eines systematischen Fehlers (Bias) bei der Schätzung der Modellgüte beschrieben, der bei der nichtvalidierten Nutzung eines Regressionsmodells für die Vorhersage von Messwerten auftritt. Dieser Vorhersagefehler soll durch die anschließend beschriebenen Validierungsmethoden erkannt und reduziert, im optimalen Fall korrigiert werden.

Definition der Modellgüte

Die am häufigsten genutzten Regressionsmodelle sind die der linearen und der logistischen Regression, je nach Merkmalstyp der interessierenden Zielgröße.

In der linearen Regression [5] wird eine stetige Zielgröße modelliert. In diesem Modell wird üblicherweise das lineare Bestimmtheitsmaß r^2 als Gütemaß für die Modellanpassung an die vorliegenden Daten genutzt. Dabei gibt r^2 an, wie viel der Variabilität der Zielgröße durch das Modell – sprich die darin enthaltenen Variablen – erklärt wird. Dieses Maß wird bestimmt, indem man die Daten der vorliegenden Beobachtungen in das Modell einsetzt und den so berechneten Modellwert mit dem vorliegenden beobachteten Messwert vergleicht. Je geringer diese Abstände sind, desto besser passt das Modell zu den Daten und hat so eine entsprechende Güte.

Die logistische Regression [7] wird eingesetzt, wenn die interessierende Zielgröße nur zwei Ausprägungen (z. B. 0/1, ja/nein, geheilt/krank) besitzt und geprüft werden soll, welche Einflussgrößen in Zusammenhang stehen mit einer solchen Zielgröße. Dabei wird durch Einsetzen der Variablen der Beobachtungen die Wahrscheinlichkeit für das interessierende Ereignis (Ausprägungen der Zielgröße) modelliert. Dieser Modellwert wird in Bezug gesetzt zur beobachteten Zielgröße, indem ein Grenzwert (Cutpoint) festgelegt wird, ab dem die Modell-Wahrscheinlichkeit für das Ereignis spricht, sonst eben nicht. Es resultiert daraus eine Vierfeldertafel, in der die so bestimmte Modell„prognose" mit der realen Beobachtung in Bezug gesetzt wird. Aus dieser Vierfeldertafel können die aus der Diagnostik bekannten Maße wie Sensitivität und Spezifität als Gütemaße berechnet werden [8]. Durch Veränderung des Cutpoints erhält man eine sogenannte ROC-Analyse (receiver operating characteristics), aus der die Fläche unter der ROC-Kurve (area under the curve – AUC) als globales Anpassungs(Güte-)maß unabhängig vom gewählten Cutpoint bestimmt wird [9].

Bias der Modellgüte (Optimismus)

Diese Gütemaße können nun, wie gerade beschrieben, für das jeweilige Regressionsmodell anhand der Daten bestimmt werden. Das Problem bei der Nutzung von solch zusammenfassenden Kennzahlen der Regressionsanalyse (z. B. r^2 und AUC) als Maß für die Performanz eines Modells (Modellgüte) ist, dass das Modell beim vorhandenen Datensatz immer eine bessere Güte hat als bei neuen Daten.

Dies liegt an dem folgenden Zirkelschluss: Zuerst werden mit den üblichen Schätzmethoden (Kleinste-Quadrate-Schätzung, Maximum-Likelihood-Schätzung) die Regressionskoeffizienten optimal für den vorliegenden Datensatz bestimmt. Anschließend wird mit dem für den vorliegenden Datensatz optimierten Modell die Güte anhand derselben Daten geschätzt. Somit ist

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Tab. 1 Übersicht über die Einteilung der gebräuchlichsten Validierungsverfahren in interne und externe Validierungsverfahren

Validierungsverfahren	intern/extern
externe Validierung	extern
Data-Splitting	intern („extern“)
Kreuzvalidierung	intern
Bootstrap-Validierung	intern
Shrinkage	intern

diese Güteschätzung [5] sicher zu optimistisch und kann als obere Schranke für die eigentliche, zugrunde liegende – aber unbekannte – Güte angesehen werden.

In der Literatur wird dieser Fehler als Vorhersagefehler oder „wahrer Fehler“ (prediction error) in den zukünftigen Stichproben bezeichnet. Der Fehler, den man bei Nutzung der vorliegenden Stichprobe, der sog. Lernstichprobe, bei der Güteschätzung durch das ermittelte Regressionsmodell erhält, wird als „scheinbarer Fehler“ (apparent error) bezeichnet [6]. Bei Nutzung einer Vorhersageregel (hier: Regressionsmodell) anhand der Daten des vorliegenden Originaldatensatzes ist die Schätzung des wahren Fehlers durch den scheinbaren Fehler meist verzerrt in Richtung zu guter Modellwerte.

Übersicht über die Validierungsverfahren

Es gibt verschiedene Validierungsverfahren. In der Praxis wird nach Bestimmung der Modellgüte auf Basis des vorhandenen Datensatzes in der Regel allerdings nur eines der Validierungsverfahren benutzt. Es gibt in Bezug auf die Nutzung der Verfahren verschiedene „Schulen“, die jeweils auf bestimmte Validierungsverfahren Wert legen. Um diese Entscheidung nicht vorzugeben, werden hier die wesentlichen in der Literatur beschriebenen und genutzten Verfahren vorgestellt. **Tab. 1** gibt einen Überblick über die vorzustellenden Verfahren und deren Einteilung als interne oder externe Validierungsmethode.

Die externe Validierung ist die weitestgehende Validierung. Dabei wird das entwickelte Regressionsmodell auf einen zweiten, unabhängigen Datensatz angewendet und die dann erreichte Güte zur Beurteilung des Modells genutzt. Die internen Validierungsverfahren beruhen auf dem Originaldatensatz, der auch für die Modellentwicklung zur Verfügung stand. Dabei werden verschiedene statistische Vorgehensweisen genutzt, um den systematischen Fehler in der Güteschätzung zu bestimmen, um diesen ggf. korrigieren zu können. Beim *Data-Splitting* wird der Originaldatensatz mit einem Zufallsverfahren in mehrere Datensätze (meist zwei) aufgeteilt. Einer dient der Modellentwicklung und der andere wird dann für eine „externe“ Validierung genutzt. *Kreuzvalidierung*[1] und *Bootstrap-Validierung* beruhen auf rechenintensiven Resamplingverfahren [10]. Dabei werden durch Stichprobenziehungen mit und ohne Zurücklegen „neue“ Datensätze erzeugt, die zur Validierung herangezogen werden. Shrinkage-Faktoren berücksichtigen die Schätzungsgenauigkeit des Regressionsmodells durch eine Korrektur der Regressionskoeffizienten. Auch dadurch wird versucht, den Optimismus bei der Schätzung der Modellgüte zu korrigieren. Dabei bleibt die Modellstruktur unangetastet.

In den nächsten Abschnitten werden die einzelnen Vorgehensweisen kurz charakterisiert. Es werden das allgemeine Vorgehen, Vor- und Nachteile sowie Spezifika der Verfahren soweit präsentiert, dass eine Entscheidung für das eine oder andere Verfahren getroffen werden kann. Für eine detaillierte Beschreibung der Verfahren muss auf die jeweils angegebene Literatur verwiesen werden.

Externe Validierung

Wie im vorigen Abschnitt beschrieben, ist bei der externen Validierung eines Regressionsmodells die Anwendung eines zweiten, unabhängigen Datensatzes gefordert. Dadurch erhält man eine Abschätzung des Fehlers, der durch die Nutzung des Originaldatensatzes für Modellentwicklung und Modellgütebestimmung eingetreten ist. Altman und Royston [4] unterscheiden in ihrem grundlegenden Artikel zwischen einer Validierung an einem zweiten Datensatz aus derselben Quelle wie der Originaldatensatz und zwischen der Validierung an einem Datensatz aus völlig anderer Quelle. Sie schreiben dazu, dass die Generalisierbarkeit und die Übertragbarkeit der Ergebnisse eigentlich nur mit der zweiten Variante geprüft werden kann. Die externe Validierung ist genau die Validierung, die zur untersuchten Fragestellung führt: Wie gut ist ein Modell bei Anwendung auf neue Daten?

Dementsprechend ist dieses Vorgehen, wann immer möglich, allen anderen Validierungsmethoden vorzuziehen. Man untersucht bei der Durchführung einer externen Validierung die Übertragbarkeit der Ergebnisse auf andere Datensätze (z.B. aus anderen Kliniken), die nicht immer gegeben sein muss. Ein gutes Beispiel für eine externe Validierung liefert der Artikel von Wagner et al. [11], in dem der APACHE-Score aus der Notfallmedizin validiert wurde.

Der Hauptnachteil der externen Validierung ist die Notwendigkeit, einen zweiten, strukturgleichen Datensatz zur Verfügung zu haben. Sehr oft wird schon aus Kosten- und Aufwandsgründen dieser zweite Datensatz nicht zu erhalten sein. Anders als im späteren Anwendungsfall der Modellierung einer klinischen Prognose muss hier in der Studiensituation die Zielgröße bereits bekannt sein, um die Modellgüte anhand dieses zweiten Datensatzes bestimmen zu können.

Data-Splitting

Durch die zufällige Aufteilung eines vorliegenden Datensatzes in zwei Teildatensätze kann eine Validierung ermöglicht werden. Diese ist, da die Daten aus einer Quelle stammen, eigentlich als internes Validierungsverfahren zu werten. Da durch diese Vorgehensweise aber zwei unabhängige Datensätze erzeugt werden, kann die Validierung auch als (quasi-)extern angesehen werden. Ein Teildatensatz wird dann für die bisher beschriebene Modellentwicklung inklusive Reklassifikation und Bestimmung der Modellgüte genutzt, der zweite Teildatensatz steht für eine „externe“ Validierung (allerdings anhand von Daten aus derselben Quelle) zur Verfügung. Die Aufteilung wird mit einem Zufallsmechanismus gesteuert, der den Originaldatensatz auf der Basis eines vorgegebenen Prozentsatzes teilt. Gängige Prozentsätze für den Anteil des Trainingsdatensatzes sind 50%, 66% oder 70%. Der jeweilige Restdatensatz bildet dann den Testdatensatz.

Die Vorteile des Data-Splitting sind erstens, dass der Testdatensatz dieselbe Struktur wie der Trainingsdatensatz hat, und zwei-

[1] In diesem Artikel wird der Begriff Kreuzvalidierung als spezielle Methode für die interne Validierung (s. Abschnitt „Kreuzvalidierung“ weiter unten) verwendet. Dieser Begriff wird in vielen Anwendungen aber auch allgemein für die Überprüfung eines Modells an neuen Daten verwendet, unabhängig von der konkreten Validierungsmethode.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

tens, dass der Testdatensatz aus derselben Grundgesamtheit wie der Modellierungsdatensatz stammt, da er durch eine Zufallsstichprobenziehung aus dem Originaldatensatz gewonnen wurde. Somit kann eine Validierung problemlos erfolgen. Leider überwiegen bei der Anwendung dieses Validierungsverfahrens die Nachteile. Unter anderem beschreiben Harrell [5], Altman und Royston [4] sowie Efron und Tibshirani [6] das Data-Splitting zur Modellvalidierung genauer und formulieren folgende Probleme:

- Data-Splitting reduziert die Fallzahl in den Datensätzen (sowohl im Trainings- als auch im Testdatensatz). Damit geht die bessere Schätzung der Modellgüte zu Lasten der Modellentwicklung (größere Streuung der Regressionskoeffizienten bis hin zur Unmöglichkeit einer Modellierung).
- Die Schätzer der Güte (z. B. r^2 oder AUC) werden ungenauer.
- Das Ergebnis hängt von der konkreten Zufallsstichprobe ab.
- Mit der Methode des Data-Splitting kann nicht ein endgültiges Modell auf Basis aller Daten validiert werden, sondern nur dasjenige, welches auf einem Teildatensatz entwickelt wurde.
- Data-Splitting erfordert, dass die Aufteilung vor Beginn der gesamten Analyse vorgenommen wird.

Kreuzvalidierung

Die Kreuzvalidierung (engl.: cross-validation) ist eine Erweiterung/Verallgemeinerung des Data-Splitting, um dessen Nachteil der Fallzahlreduzierung zu umgehen. Dabei wird der Datensatz mehrfach in K Gruppen unterteilt und die Auswertungen (Modellentwicklung und -validierung) werden jeweils in diesen Untergruppen durchgeführt [5].

Die Durchführung der sog. „K-fold cross-validation“ basiert auf einer Stichprobenziehung ohne Zurücklegen [10]: Dabei wird der Datensatz mit einem Zufallsverfahren in K Untergruppen aufgeteilt. Eine dieser Untergruppen wird anschließend aus dem Datensatz herausgenommen und fungiert als Testdatensatz. Der Datensatz der restlichen K-1 Untergruppen ist dabei der Trainingsdatensatz, an dem die Regressionskoeffizienten des Modells bestimmt werden. Dieses Modell wird auf den Testdatensatz angewendet, und es wird jeweils die Güte des Modells bestimmt. Dieses Vorgehen wird für alle K Gruppen durchgeführt, so dass anschließend jeweils K Kenngrößen vorliegen, die durch Mittelwertbildung zusammengefasst werden und die kreuzvalidierten Gütekenngrößen darstellen.

Diese Auswertung hängt jedoch von der konkreten Zufallsziehung in die K Untergruppen ab. Um diese Abhängigkeit zu verringern, wird die gesamte Auswertung mit jeweils anderer Zufallsziehung X-mal wiederholt. Typische Werte für X bei der Kreuzvalidierung sind 20–50 Wiederholungen [5]. Die Zusammenfassung über die X·K Auswertungen ergibt dann den kreuzvalidierten Schätzer für die Güte.

Für die Auswahl der Anzahl der Untergruppen (K) geben Davison und Hinkley [12] die Hilfsformel an, dass die Anzahl der Gruppen entweder 10 nicht übersteigen sollte, um nicht zu viele Modelle rechnen zu müssen, oder (n = Fallzahl), um eine gewisse Anzahl Beobachtungen in den Untergruppen zu garantieren, und empfehlen den kleineren der beiden Werte zu nutzen.

Davison und Hinkley geben noch eine adjustierte Variante für den kreuzvalidierten Schätzer an ([12], S. 295). Der adjustierte Schätzer hat einen kleineren Bias und ist nur mit geringfügig höherem Aufwand zu berechnen. Eine extreme Form der Kreuzvalidierung wird als „Jackknife“ oder auch „leaving-one-out“ bezeichnet. Beim Jackknife-Vorgehen wird K = n gesetzt, d. h., dass jede Beobachtung im Datensatz eine Untergruppe bildet. Es wird also jeweils eine Beobachtung eliminiert, mit den restlichen die Modellierung durchgeführt und auf Basis dieses Modells und der Messwerte dieser Beobachtung die Prognose geschätzt. So erhält man für jede Beobachtung eine Prognose und den beobachteten Wert und kann so die Prognosegüte bestimmen. Bei der Wahl von K = n beim Jackknife ist die Gefahr eines relevanten Bias in der Schätzung der Modellgüte aber recht groß, so dass eine Validierung mit dem Jackknife-Schätzer nicht allgemein empfohlen werden kann.

Folgende Probleme werden in der Literatur (z. B. Harrell [5], S. 93–94) bei der Nutzung der Kreuzvalidierung genannt:

- Wie beim Data-Splitting werden die Modellschätzungen nicht auf der vollen Information der n Beobachtungen durchgeführt. Der Effekt ist nicht groß, aber offensichtlich. Dadurch wird vom Grundsatz her nicht unbedingt das Modell, welches auf der Basis von n Beobachtungen entwickelt wurde, validiert.
- Die Auswertung hängt von der Wahl von K ab, so dass eine gewisse Subjektivität in die Ergebnisse gelangen kann.
- Die Anzahl notwendiger Wiederholungen (>20) führt zu sehr vielen Modellschätzungen.

Als allgemeine Hintergrundliteratur zur Schätzung der Vorhersagefehler durch Kreuzvalidierung sind das Kapitel 17 der Monografie von Efron und Tibshirani [6] und das Kapitel 5.2 des Buches von Harrell [5] zu empfehlen.

Bootstrap-Validierung

Der Begriff Bootstrap leitet sich von der Redewendung „to pull oneself up by one‘s bootstrap“ (sich selbst an den eigenen Stiefelschlaufen hochziehen) ab und spielt auf die „Abenteuer des Barons von Münchhausen“ von Rudolf Erich Raspe an. Hinter der Idee des Bootstraps verbirgt sich ein Ziehen mit Zurücklegen: Aus einem Datensatz wird durch zufälliges Ziehen mit Zurücklegen ein neuer Datensatz gleicher Fallzahl gezogen [6, 10, 12]. Dabei wird jeweils eine Beobachtung gezogen und dem neuen Datensatz zugefügt, ohne im Ausgangsdatensatz gelöscht zu werden. Beim nächsten Zug kann zufällig dieselbe Beobachtung noch einmal gezogen werden. Es werden so viele Ziehungen durchgeführt, wie Beobachtungen im Ausgangsdatensatz sind. Dieser so erzeugte „neue“ Datensatz wird als Bootstrap-Sample bezeichnet. Dadurch erhält man einen dem Ausgangsdatensatz ähnlichen Datensatz mit gleichen statistischen Eigenschaften, in dem die Beobachtungen des Ausgangsdatensatzes keinmal, einmal oder mehrmals vorkommen können. Die Idee ist nun, dass durch mehrfache Erzeugung solcher Bootstrap-Samples und Wiederholung der Auswertungen viele Gütemaße erzeugt werden und man die Verteilung dieser Maße untersuchen kann. Unter anderem lassen sich durch Mittelwertbildung ein Schätzwert für den wahren, aber unbekannten Wert erzeugen und auch Konfidenzintervalle dafür berechnen.

Für die Modellvalidierung auf Grundlage einer Regression bedeutet dies nun, anhand vieler Bootstrap-Samples die Modellgüten zu bestimmen und daraus den validierten Schätzer zu ermitteln. Efron und Tibshirani empfehlen mindestens 50–200 Bootstrap-Samples zur Schätzung des Standardfehlers und mindestens 1 000 Bootstrap-Samples zur Schätzung von Konfidenzintervallen.

Mit dem Bootstrap-Vorgehen wurde von Efron und Tibshirani [6] ein Vorgehen beschrieben, bei dem keine Beobachtungen zur Schätzung zurückgehalten werden müssen, wie das beim Data-Splitting oder bei der Kreuzvalidierung notwendig ist. Dabei

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

wird meist vom intuitiven Ansatz des „simple bootstrap" ausgegangen, bei dem in jedem Bootstrap-Sample die Regressionskoeffizienten neu geschätzt werden und dieses so erzeugte Modell auf den Originaldatensatz zur Schätzung der Güte angewendet wird. Durch eine Mittelung dieser meist mehrere hundert Schätzer ergibt sich dann der Simple-Bootstrap-Schätzer nach Modellvalidierung.

Efron und Tibshirani zeigen, dass eine etwas komplexere Variante, der „enhanced bootstrap"-Schätzer, bessere Eigenschaften bezüglich des Bias hat. Bei dieser Variante wird der Schätzer nicht direkt durch Mittelwertbildung der Bootstrap-Ergebnisse bestimmt. Vielmehr schätzt man hierbei den Optimismus, den Schätzer für den gesuchten Bias. Nach Schätzung des Optimismus kann anschließend durch Subtraktion von der Güte des Originalmodells der Enhanced-Bootstrap-Schätzer ermittelt werden. Das allgemeine Vorgehen zur Durchführung des Enhanced Bootstrap kann folgendermaßen zusammengefasst werden:

1. Berechne die Modellgüte am Originaldatensatz,
2. erzeuge X Bootstrap-Samples durch Ziehen mit Zurücklegen aus dem Originaldatensatz,
3. passe in jedem Bootstrap-Sample das Modell neu an (Neuberechnung der Regressionskoeffizienten, sogenanntes Bootstrap-Modell),
4. wende jedes Bootstrap-Modell auf den Originaldatensatz an und berechne jeweils die Modellgüte, der Mittelwert über diese X Gütemaße ergibt den *Simple-Bootstrap-Schätzer.*
5. Wende jedes Bootstrap-Modell auch auf das jeweilige Bootstrap-Sample an und berechne jeweils die Modellgüte.
6. Der Optimismus wird dann geschätzt als der Mittelwert der Differenzen der beiden Gütekriterien eines jeden Bootstrap-Modells bei Anwendung auf den Originaldatensatz und auf das Bootstrap-Sample.
7. Der *Enhanced-Bootstrap*-Schätzer ist dann die Güte des Modells am Originaldatensatz nach Reklassifikation (siehe 1.), reduziert um den Optimismus (siehe 6.).

Die Vor- und Nachteile des Bootstraps können folgendermaßen zusammengefasst werden:

- Der Hauptvorteil dieses Verfahrens gegenüber den anderen Validierungsverfahren liegt an der relativ stabilen und varianzarmen Schätzung der Modellparameter [5].
- Beim Bootstrap wird die Validierung auf dem gesamten Datensatz (n Beobachtungen) durchgeführt. Es müssen nicht wie beim Data-Splitting und bei der Kreuzvalidierung Beobachtungen bei der Wiederholung der Modellanpassung herausgenommen werden.
- Bootstrap kann auf jede beliebige einfache oder komplizierte Statistik angewendet werden.
- Es müssen keine Verteilungsannahmen gemacht werden, da die Kenngrößen aus der empirischen Verteilung geschätzt werden.
- Die Auswertung muss automatisierbar sein, da sie hundertfach ablaufen soll.
- Die Güte der Bootstrap-Schätzer hängt von der Datenqualität der zugrunde liegenden Stichprobe ab.
- Bei ausreißerempfindlichen Statistiken kann das Ergebnis verfälscht werden, wenn einige Ausreißer im Originaldatensatz vorhanden sind.
- Ein Kritikpunkt ist die unter Umständen hohe Rechenzeit bei großen Datensätzen.

Als Übersichtsarbeiten zur Bootstrap-Validierung eignen sich die Lehrbücher von Efron und Tibshirani [6] und Davison und Hinkley [12], in denen neben der allgemeinen Herleitung des Vorgehens beim Bootstrap auch die Anwendung auf den Vorhersagefehler in Regressionsmodellen enthalten ist.

Shrinkage-Faktoren

Seit langem sind Verfahren bekannt, Schätzwerte für Parameter in linearen Modellen, die nach der Methode der kleinsten Quadrate bzw. nach der Maximum-Likelihood-Methode gewonnen wurden, mit einem Faktor c ($0<c\leq 1$), dem sog. Shrinkage-Faktor, zu multiplizieren und in diesem Sinne zu verkürzen, um eine „bessere" Anpassung zu erreichen [13]. Dadurch wird die Erwartungstreue der Schätzer im Allgemeinen nicht mehr gewährleistet sein, der Mean-Square-Error kann jedoch verringert werden.

Van Houwelingen und Le Cessie geben in ihrer Übersichtsarbeit zur Shrinkage-Methode [13] allgemeine Schätzmöglichkeiten für Shrinkage-Faktoren im linearen, logistischen und Cox-Modell an. Die Idee ist die Kalibrierung (Korrektur) der Regressionskoeffizienten, um den Optimismus im Modell zu korrigieren. Damit bekommt man am Ende ein korrigiertes Modell, das intern validierte Kenngrößen liefert.

Van Houwelingen und Le Cessie beschreiben einen relativ einfachen, aus der Regression nach Maximum-Likelihood-Schätzung zu berechnenden heuristischen Shrinkage-Faktor c, da die zur Berechnung benötigten Werte direkt aus den meisten Statistiksoftware-Paketen zu erhalten sind und der heuristische Shrinkage-Faktor somit ohne großen Aufwand geschätzt werden kann. Die Eigenschaften von c können folgendermaßen beschrieben werden: Im Allgemeinen ist $c<1$. Wenn c nahe bei 0 liegt, hat der Regressionskoeffizient keinen Einfluss auf die Zielgröße, wenn er nahe 1 ist, gibt es kaum Korrektur und somit nur wenig Optimismus.

Sauerbrei [14] beschreibt zu Recht, dass ein globaler Faktor zur Kalibrierung aller Regressionskoeffizienten sicher nicht optimal sein kann. Er hat deshalb in seiner Arbeit einen Vorschlag vorgelegt, wie man jeweils einen speziellen Korrekturfaktor für jeden Koeffizienten erhalten kann. Dabei nutzt er nicht die auf Basis der Modelle geschätzten Werte, sondern die entsprechenden Modellanteile mit den jeweiligen Regressionskoeffizienten. Allerdings ist es bei diesem Ansatz möglich, für einzelne Regressionskoeffizienten Shrinkage-Faktoren <0 oder >1 zu erhalten. Sauerbrei empfiehlt, negative Faktoren auf 0 zu setzen. Faktoren >1 sollen wie berechnet angewendet werden, da es bei diesem Ansatz vorkommen kann, dass der Einfluss eines einzelnen Faktors nicht über-, sondern unterschätzt ist.

Vergleich der internen Validierungsverfahren

Es gibt in der Literatur einige wenige Untersuchungen zum Verhalten der internen Validierungsverfahren in Regressionsmodellen, die hier zusammenfassend dargestellt werden sollen.

Steyerberg et al. ([15], S. 780) untersuchen das Verhalten verschiedener interner Validierungsmethoden bei Anwendung der logistischen Regression. In einer Simulationsstudie werden dabei das Data-Splitting, K-fold-Kreuzvalidierung und Enhanced Bootstrap in Abhängigkeit von der Fallzahl verglichen. Die Autoren kommen in der Zusammenfassung zu dem Schluss: „Optimal estimates of internal validity of logistic regression models constructed in smaller samples … are generally obtained with the regular variant of bootstrapping" (mit „regular" ist hier die Enhanced-Bootstrap-Variante gemeint).

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Kenngröße	ohne Validierung	10-fold CV (50)	Jackknife	Enhanced Bootstrap	heuristischer Shrinkage	Parameterwise Shrinkage
AUC	0,88	0,84	0,88	0,86	0,88	0,88
Sensitivität*	0,82	0,89	0,81	0,77	0,84	0,80
Spezifität*	0,78	0,75	0,78	0,78	0,76	0,78
Youden-Index*	0,61	0,63	0,59	0,55	0,60	0,59
PPV**	0,40	0,41	0,40	0,36	0,38	0,40
NPV**	0,96	0,98	0,96	0,96	0,96	0,96

* am Cutpoint mit maximalem Youden-Index, ** und bei Nutzung der empirischen Prävalenz von 15 %, CV = Cross-Validation, AUC = Area under the Curve, PPV = positiver prädiktiver Wert, NPV = negativer prädiktiver Wert

Tab. 2 Ergebniszusammenfassung der Modellvalidierung im Beispiel (aus [18])

Harrell betont ([5], S. 96), dass die Bootstrap-Schätzer eher zu Gunsten des Modells verzerrt sein können. Die Kreuzvalidierung ist weniger verzerrt als der Bootstrap, hat aber üblicherweise eine größere Varianz der Schätzer und ist demnach wesentlich ungenauer. Dies bedeutet, dass die Kreuzvalidierung bei Wiederholung des gesamten Modellierungsprozesses andere Ergebnisse liefern kann.

Harrell vergleicht in einer weiteren Untersuchung [16] einige Validierungsmethoden. Die Simulationen wurden mit einer Fallzahl von n = 200 bei einer Variablenzahl von 15 bzw. 30 durchgeführt. Hier kommt er zu dem Schluss, dass die Bootstrap-Validierung mindestens gleich gut oder besser als die Kreuzvalidierung ist.

Davison und Hinkley ([12], S. 362) fassen zusammen: „It seems likely, in the light of the preceding example, that the bootstrap approach would be preferable."

Sauerbrei [14] zeigt, dass der hier erstmals vorgestellte Parameterwise-Shrinkage-Ansatz Vorteile gegenüber dem heuristischen Shrinkage bietet. Außerdem wird eine Parallele zwischen Jackknife-Schätzer und heuristischem Shrinkage dargestellt. Einige Simulationsergebnisse zum Verhalten gegenüber dem Bias favorisieren den Bootstrap-Ansatz.

Zusammenfassend kann man festhalten, dass für den Einsatz bei der Modellvalidierung von Regressionsmodellen in den meisten Fällen der Enhanced-Bootstrap-Schätzer als geeigneter Schätzer für die Güte angesehen wird. Aufgrund der methodischen Probleme sind das Data-Splitting, der Jackknife-Schätzer und die globalen Shrinkage-Faktoren für den allgemeinen Gebrauch wohl nicht zu empfehlen.

Anwendungsbeispiel aus der Rehabilitationsforschung

Zur Beurteilung der Effektivität von Rehabilitationsmaßnahmen spielt vor allem das Behandlungsergebnis eine wichtige Rolle. Ein wesentlicher Outcome ist dabei langfristig der Erhalt der Erwerbsfähigkeit. Für eine effektive Rehabilitation ist es deshalb wichtig, jene Patienten möglichst früh zu erkennen, denen aller Wahrscheinlichkeit nach eine spätere Erwerbsunfähigkeit droht, um z. B. unterstützende Maßnahmen bereits während der Rehabilitationsmaßnahme oder in einer entsprechenden Nachsorge durchführen zu können. Als Beispiel in dieser Arbeit dient die im Rehabilitationswissenschaftlichen Verbund Ulm durchgeführte Untersuchung, wie gut mit einem Prognosemodell mittelfristig (1–2 Jahre) nach einer stationären Rehabilitationsmaßnahme eine Erwerbsunfähigkeit prognostiziert werden kann [7, 17]. Bei Patienten mit hoher Wahrscheinlichkeit einer drohenden Erwerbsunfähigkeit könnten also konsequentere Maßnahmen getroffen werden, die dies verhindern könnten.

Die statistische Auswertung schloss 841 behandelte Patienten ein, die ein stationäres Heilverfahren durchführten. Spezielle Ausschlusskriterien gab es nicht, um ein möglichst repräsentatives Kollektiv zu erhalten. Insgesamt wurden 131 (15,6 %) Erwerbsunfähigkeitsrentner und 710 (84,4 %) Erwerbstätige beobachtet. Die Auswertung wurde für die Zielgröße Erwerbsunfähigkeitsrente ja = 1, nein = 0 mit dem logistischen Regressionsmodell ausgewertet und beispielhaft in Muche [7] und Muche et al. [18] ausführlich beschrieben.

In diesem Modell zeigen die Variablen Alter, Arbeitsunfähigkeit und Leistungsfähigkeit des Patienten den größten prognostischen Effekt auf die Erwerbsunfähigkeitsrente. Weitere sieben Routinevariable konnten als grenzwertig einflussreich identifiziert werden. Eine ROC-Analyse der prognostischen Wertigkeit des Modells wurde anschließend durchgeführt, um zu untersuchen, wie gut sich die Zielgröße aus dem Modell vorhersagen lässt. Dafür werden die üblichen Kenngrößen für Prognosen (Sensitivität, Spezifität, prädiktive Werte und die Area under the Curve [AUC] als globales Prognosemaß) bestimmt, die gut nachvollziehbar in Faller [8] in dieser Serie dargestellt werden sowie mit mehr Hintergrund in Pepe [9] nachzulesen sind. Deshalb wird in diesem Zusammenhang auch die Modellgüte als Prognosegüte bezeichnet.

Die Anwendung auf die vorliegenden Daten ergibt mit einem AUC-Wert von 0,88 eine sehr gute Prognosegüte für dieses Modell (AUC >80 % werden allgemein als sehr gut, AUC >90 % als exzellent eingestuft [9]). Auch die weiteren Prognosegütemaße sind mit einer Sensitivität von 0,82 und einer Spezifität von 0,78 relativ gut für die Prognose einer Erwerbsunfähigkeitsrente in der Situation, dass nur Routinevariablen für diese Prognose genutzt werden. Dieses Modell wurde dann mit den hier beschriebenen Validierungsverfahren auf einen Optimismus in der Prognosegüte untersucht. Die resultierenden Ergebnisse werden hier zusammenfassend dargestellt (**Tab. 2**). Dabei werden Sensitivität, Spezifität und prädiktive Werte (positive predictive value [PPV], negative predictive value [NPV]) jeweils bei dem Cutpoint der stetigen Wahrscheinlichkeiten für eine Erwerbsunfähigkeit angegeben, bei der der Youden-Index als Summe von Sensitivität und Spezifität (genauer: Sensitivität + Spezifität –1) am größten wird. Dieses Kriterium der Maximierung des Youden-Index bedeutet, dass man ein gemeinsames Maß der beiden Kenngrößen Sensitivität und Spezifität benutzt und dass beide als gleichwertig betrachtet werden. Der Cutpoint ist hier ein Wert zwischen 0 und 1, an dem die Wahrscheinlichkeit für ein Ereignis in Prognose für Erwerbsunfähigkeit (Wahrscheinlichkeit größer als Cutpoint) und Prognose für Erwerbsfähigkeit (Wahrscheinlichkeit kleiner Cutpoint) aufgeteilt wird, um die Sensitivität und Spezifität berechnen zu können. Für die Angabe der prädiktiven Werte ist außerdem noch eine Prävalenz anzu-

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

geben. Diese wurde hier mit der empirischen Prävalenz der Erwerbsunfähigkeiten (15%) angenommen.

Legt man nach den obigen Empfehlungen die Ergebnisse der Enhanced-Bootstrap-Validierung für ein endgültiges Fazit zum vorliegenden Prognosemodell zugrunde, dann erkennt man nur relativ kleine Korrekturen der Prognosekenngrößen bezüglich des Optimismus. Die AUC verringert sich um etwa 2 Prozentpunkte gegenüber dem unvalidierten Ergebnis, bleibt mit einem Wert von etwa 86% für Prognosen in einem akzeptablen Bereich. Bei den Cutpoint-abhängigen Maßen werden die Sensitivität (von 0,82 auf 0,77) und der positive prädiktive Wert (von 0,40 auf 0,36) am meisten nach unten korrigiert, allerdings wird die Gesamtaussage zur Nutzung des Modells dadurch nicht in Frage gestellt. Das in dem Projekt aufgestellte logistische Regressionsmodell für die Prognose einer Erwerbsunfähigkeit nach stationärer Rehabilitation zeigt somit wenig Tendenz, zu optimistisch zu sein.

Beim Vergleich der Validierungsverfahren zeigen sich die erwarteten Ergebnisse. Das Jackknife-Verfahren sowie der heuristische Shrinkage sind einander sehr ähnlich und korrigieren die Prognosegütemaße des Originalmodells kaum. Auch die Kreuzvalidierung ergibt sehr ähnliche Ergebnisse. Der Parameterwise Shrinkage verändert die Prognosegüte kaum. Die größten Korrekturen zeigen sich bei der Kreuzvalidierung und dem Enhanced Bootstrap, wobei die Ergebnisse des Enhanced Bootstrap mit den zu erwartenden Ergebnissen konsistenter sind. So wird z.B. bei der Kreuzvalidierung das zusammenfassende Maß bei den Cutpoint-abhängigen Maßen, der Youden-Index, größer geschätzt als vor der Validierung. Betrachtet man alle Ergebnisse der verschiedenen Validierungsverfahren und die Erkenntnisse aus den vergleichenden Untersuchungen im letzten Abschnitt, so kommt man zu dem Schluss, dem Ergebnis des Enhanced Bootstrap am meisten zu trauen, und man würde diese Werte als Ergebnis der internen Validierung für das Prognosemodell angeben.

Insgesamt betrachtet kann man etwas Optimismus erkennen. Die für die Prognose genutzten Routinevariablen zeigen ein Potenzial von 77–78% Sensitivität und Spezifität, mögliche Erwerbsunfähigkeitsrentner zu Beginn einer stationären Rehabilitationsmaßnahme zu erkennen. Die prädiktiven Werte (eher niedriger positiver prädiktiver Wert, sehr hoher negativer prädiktiver Wert) zeigen, dass das Prognosemodell eher als Ausschlusstest genutzt werden sollte, da die Prognose für „keine Erwerbsunfähigkeitsrente" sehr genau gestellt werden kann. Für die „prognosepositiven" Patienten sollten weitere Untersuchungen durchgeführt bzw. weitere Informationen erhoben werden, um auch hier genauere Vorhersagen treffen zu können.

Fazit

Zusammenfassend kann man folgende mögliche Konsequenzen einer Modellvalidierung angeben:

- Im Falle, dass die Güte des Modells schon vor einer Modellvalidierung für einen Einsatz im Alltag zu gering ist, ist eine Validierung nicht notwendigerweise durchzuführen. Durch die Validierung erfolgt hauptsächlich eine Reduzierung der Güte, und so fallen die dann korrigierten Maße noch schlechter aus. Ein Modell, was sich von vornherein als nicht brauchbar herausstellt, muss nicht bis ins Detail untersucht und validiert werden.
- Will man jedoch das Modell in der Praxis einsetzen und anschließend mit Hilfe der so geschätzten Werte z.B. Therapieentscheidungen treffen, sollte die Performanz des Modells für neue Daten untersucht worden sein. Die Entscheidung für die Brauchbarkeit in der Praxis sollte immer erst nach der Validierung getroffen werden.
- Besonders wenn der Datensatz eine relativ kleine Fallzahl besitzt oder die Streuung der Regressionskoeffizienten relativ groß ist, können sich große Unterschiede ergeben.

In Bezug auf die Methodik kann man feststellen, dass für den Einsatz bei der Modellierung in den meisten Fällen der Enhanced-Bootstrap-Schätzer als geeigneter Schätzer für die Güte nach Modellvalidierung angesehen wird. Aufgrund der methodischen Probleme sind das Data-Splitting, der Jackknife-Schätzer und die globalen Shrinkage-Faktoren wohl für den allgemeinen Gebrauch nicht zu empfehlen.

Somit hat man mit den internen Validierungsmethoden Vorgehensweisen zur Hand, die eine Abschätzung des Vorhersagefehlers ohne Vorliegen eines zweiten, unabhängigen Datensatzes ermöglichen. Aber alle zitierten Autoren, die sich mit Modellvalidierung beschäftigt haben, beschreiben, dass eine interne Validierung zwar ein probates Mittel zur Abschätzung des Optimismus und zur Korrektur der Modellgüte sein kann, dass jedoch das endgültige Prognosemodell später möglichst doch noch extern validiert werden sollte. Im externen Datensatz

- gibt es u.U. eine andere Inzidenz bzw. Prävalenz der Zielgröße,
- können die Effekte einiger Einflussgrößen anders sein (Konsequenz: anderes Modell),
- gibt es u.U. eine etwas andere Patientenzusammensetzung (Ein-/Ausschlusskriterien),
- können die Ziel- und Einflussgrößen anders definiert oder beobachtet worden sein.

In der Konsequenz zeigt in der Regel die interne Validierung immer noch eine höhere Modellgüte als eine externe Validierung und lässt das Regressionsmodell zu schnell für die Praxis als geeignet erscheinen.

Danksagung

Zum Abschluss soll den Koordinatoren der Methodenreihe in der Zeitschrift „Die Rehabilitation", Hermann Faller, Thomas Kohlmann und Christian Zwingmann für ihre Anmerkungen und guten Verbesserungsvorschläge sehr herzlich gedankt werden.

Literatur

1 *Wirtz M, Morfeld M, Igl W, Kutschmann M, Leonhart R, Muche R, Schön G.* Organisation methodischer Beratung und projektübergreifender Forschungsaktivitäten in multizentrischen Forschungsprogrammen – Erfahrungen der Methodenzentren im Verbundforschungsprogramm Rehabilitationswissenschaften. Rehabilitation 2007; 46: 145–154
2 *Wyatt JC, Altman DG.* Commentary: Prognostic models: clinically useful or quickly forgotten? Brit Med J 1995; 311: 1539–1541
3 Deutsche Gesellschaft für Transfusionsmedizin und Immunhämatologie – DGTI. Computer-Validierung. Ein Leitfaden für die Validierung computergestützter Systeme bei Blutbanken. 3. Aufl. 2003, – verfügbar unter: http://www.dgti.de/downloads/downloads/Val_Leitfaden_Auflage3_TeilA.doc (aufgerufen 8.2.2008)
4 *Altman DG, Royston P.* What do we mean by validating a prognostic model? Statist Med 2000; 19: 453–473
5 *Harrell Jr FE.* Regression modeling strategies. New York: Springer 2001

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

6 *Efron B*, *Tibshirani RJ*. An introduction to the bootstrap. New York: Chapman & Hall 1993
7 *Muche R*. Die logistische Regression – ein vielseitiges Analyseinstrument rehabilitationswissenschaftlicher Forschung. Rehabilitation 2008; 47: 56–62
8 *Faller H*. Sensitivität, Spezifität, positiver und negativer Vorhersagewert. Rehabilitation 2005; 44: 44–49
9 *Pepe MS*. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press 2004
10 *Schumacher M*, *Holländer N*, *Sauerbrei W*. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? Statist Med 1997; 16: 2813–2827
11 *Wagner DP*, *Knaus WA*, *Draper EA*. Statistical validation of a severity of illness measure. Am J Public Health 1983; 73: 878–884
12 *Davison AC*, *Hinkley DV*. Bootstrap methods and their application. Cambridge: Cambridge University Press 1997
13 *Houwelingen H van*, *Le Cessie S*. Predictive value of statistical models. Statist Med 1990; 9: 1303–1325
14 *Sauerbrei W*. The use of resampling methods to simplify regression models in medical statistics. Appl Statist 1999; 48: 313–329
15 *Steyerberg EW*, *Harrell FE*, *Borsboom GJJM*, *Eijkemans MJC*, *Vergouwe Y*, *Habbema JDF*. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54: 774–781
16 *Harrell* Jr *FE*. Comparison of strategies for validating binary logistic regression models 1998, – verfügbar unter: http://biostat.mc.vanderbilt.edu/twiki/pub/Main/RmS/logistic.val.pdf (aufgerufen am 8.2.2008)
17 *Muche R*, *Rösch M*, *Flierl S*, *Alt B*, *Jacobi E*, *Gaus W*. Entwicklung und Validierung eines Prognosemodells zur Vorhersage der Arbeitsfähigkeit nach Rehabilitation anhand routinemäßig erhobener Parameter. Rehabilitation 2000; 39: 262–267
18 *Muche R*, *Ring C*, *Ziegler C*. Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Aachen: Shaker 2005

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.