

Andrej König

Der Nutzen standardisierter Risikoprognoseinstrumente für Einzelfallentscheidungen in der forensischen Praxis

Standardisierte Risikoprognoseinstrumente erfreuen sich sowohl in der Praxis als auch in der forensischen Forschung einer zunehmenden Beliebtheit. Aufgrund methodischer Probleme ist jedoch der praktische Nutzen von Risikoprognoseinstrumenten für die individuelle Vorhersage delinquenten Verhaltens fraglich.

Ziel dieser Übersichtsarbeit ist es, grundlegende methodische Schwierigkeiten – wie die Problematik der geringen Basisraten von Delinquenz, die Area under the Curve als Goldstandard in der Risikoprognoseforschung und die Heterogenität der Ausgangsstichproben – zu beschreiben und deren Auswirkungen auf Einzelfallentscheidungen kritisch zu diskutieren.

Ab welcher empirisch bestimmten Rückfallwahrscheinlichkeit ein Individuum als »hoch«, »moderat« oder »niedrig« gefährlich anzusehen ist, wird dagegen immer eine politische, juristische und ethische Frage bleiben, die sich auf dem statistischen Weg nicht beantworten lassen wird.

Schlüsselwörter: Risikoprognoseinstrumente, prädiktive Validität, Normkriterien, Rückfälligkeit

Usefulness and practicality of structured risk assessment instruments in forensic psychiatry

Using structured risk assessment instruments to predict criminal offences in individual cases has several methodological limitations. Nevertheless actuarial risk assessment instruments enjoy an increase in popularity in research and forensic practice.

This review highlights and discusses general difficulties in forensic research concerning low base rates of delinquency, the area under the curve as the gold standard in risk prediction research and the problem of heterogeneous samples, with regard to individual risk prediction.

To determine the relapse probability as »low«, »medium« or »high« will always be a political, legal and ethical question. Statistics are unlikely to resolve this problem.

Key words: Risk assessment instruments, predictive value, relapse prevention, forensic psychiatry

Die empirisch begründete Vorhersage delinquenten Verhaltens mithilfe standardisierter Risikoinstrumente wirft einige grundlegende methodische Probleme auf, die den praktischen Nutzen für den Einzelfall infrage stellen. Dessen ungeachtet erfreuen sich Risikoprognoseinstrumente sowohl in der Praxis als auch in der forensischen Forschung zunehmender Beliebtheit. Trotz fehlender Validitätsstudien finden im angloamerikanischen Raum Risikoprognoseinstrumente bereits in unterschiedlichen juristischen Kontexten (z. B. Lockerungsentscheidungen, Entziehung des Sorgerechts, Beurteilung der Schuldfähigkeit und Verurteilungen zur Todesstrafe) ihre Anwendung (WALSH & WALSH 2006). Entsprechend diesem Trend postulieren QUINSEY et al. (2003): »What we are advising is not the addition of actuarial methods to existing practice, but rather the complete replacement of existing methods with actuarial methods.« (S. 171) Die Autoren empfehlen, auf eine klinische Interpretation oder Einschätzung der Gefährlichkeit zu verzichten und die Beurteilung ausschließlich mithilfe aktueller Risikoprognoseinstrumente durchzuführen. Dagegen herrschte in den 1990er-Jahren aufgrund niedriger korrelativer Zusammenhänge zwischen klinisch-forensischen Einschätzungen und Rückfälligkeit ($r = .03-.11$) die Meinung, dass eine empirisch begründete Vorhersage gewalttätigen Verhaltens nicht möglich ist (ANDREWS & BONTA 2003). So postulierte MELOY (1992, S. 949): »It is clear from the research literature that we cannot, and will never be able to, predict

with reasonable medical certainty future violence.« Diese eher pessimistische Haltung gegenüber Risikoprognoseinstrumenten scheint sich – zumindest bei einigen Autoren – in den letzten Jahren erneut durchzusetzen (z. B. HART, MICHIE & COOKE 2007; COOKE & MICHIE 2009).

Eine umfangreiche Übersicht der international gängigsten Risikoprognoseinstrumente bieten DAHLE, SCHNEIDER und ZIETHEN (2006).

Die sogenannte Basisratenproblematik (z. B. DONALDSON & WOLLERT 2008; NEDOPIL 2007; DAHLE 2006), die »Area under the Curve (AUC)« als alleiniger Validitätskoeffizient (RICE & HARRIS 1995) und die Heterogenität der Ausgangsstichproben (DOLAN & DOYLE 2000) sind einige methodische Probleme, die in der Literatur zunehmend diskutiert werden. Ziel dieser Übersichtsarbeit ist es, die praktische Relevanz dieser methodischen Probleme bei der Anwendung von standardisierten Risikoprognoseinstrumenten zur individuellen Gefährlichkeitsprognose zu beschreiben und zu diskutieren. In Tabelle 1 werden einige grundlegende Begrifflichkeiten zusammenfassend dargestellt.

Tab. 1: Übersicht der im Text verwendeten statistischen Kennwerte

Statistischer Kennwert		Definition
Basisrate	$p(\text{rück}+)$	Anteil der rückfälligen Probanden in einer Stichprobe oder Population.
Sensitivität (Richtig-Positiv-Rate)	$p(\text{test}+ \text{rück}+)$	Anteil der als rückfällig erkannten Probanden an der Gesamtheit der rückfälligen Probanden.
Spezifität (Richtig-Negativ-Rate)	$p(\text{test}- \text{rück}-)$	Anteil der als nicht-rückfällig erkannten Probanden an der Gesamtheit der nicht-rückfälligen Probanden.
Falsch-Positiv-Rate	$p(\text{test}+ \text{rück}-)$	Anteil der als rückfällig klassifizierten Probanden an der Gesamtheit der nicht-rückfälligen Probanden.
Falsch-Negativ-Rate	$p(\text{test}- \text{rück}+)$	Anteil der als nicht-rückfällig klassifizierten Probanden an der Gesamtheit der rückfälligen Probanden.
Relevanz (Positiver-Prädiktiver-Wert)	$p(\text{rück}+ \text{test}+)$	Anteil der richtig als rückfällig erkannten Probanden an der Gesamtheit der als rückfällig klassifizierten Probanden.
Segreganz (Negativer-Prädiktiver-Wert)	$p(\text{rück}- \text{test}-)$	Anteil der richtig als nicht-rückfällig erkannten Probanden an der Gesamtheit der als nicht-rückfällig klassifizierten Probanden
Trefferquote	$p(\text{rück}+ \cap \text{test}+)$ + $p(\text{rück}- \cap \text{test}-)$	Anteil aller richtig klassifizierten Probanden.
Fehlerquote	$p(\text{rück}+ \cap \text{test}-)$ + $p(\text{rück}- \cap \text{test}+)$	Anteil aller falsch klassifizierten Probanden.
Area-Under(the)-Curve (AUC)		Das AUC gibt die Wahrscheinlichkeit an, dass ein zufällig ausgewählter rückfälliger Proband einen höheren Risikoscore hat, als ein zufällig ausgewählter nicht-rückfälliger Proband.

Bemerkung: rück+ = Rückfall; rück- = kein Rückfall; test+ = Rückfall vorhergesagt und test- = kein Rückfall vorhergesagt.

Die AUC als Validitätskoeffizient – Goldstandard oder Umgehung der Basisratenproblematik?

RICE und HARRIS (1995) führten die AUC als Validitätskoeffizienten in die forensische Risikoprognoseforschung ein. Ihr Ziel war es, einen statistischen Kennwert zu finden, der unabhängig von Basisraten einen direkten Vergleich der prädiktiven Validität verschiedener Risikoprognoseinstrumente ermöglicht. Es handelt sich dabei also um einen statistischen Versuch, ein methodisches Problem zu lösen. Die Basisraten der Rückfälligkeit variieren in Abhängigkeit der Stichprobenszusammensetzung und je nach Auswahl der Kriteriumsvariable deutlich, sodass ein empirischer Vergleich der Güte verschiedener Risikoprognoseinstrumente nicht oder nur bedingt möglich ist. In verschiedenen Studien innerhalb des deutschsprachigen Raums mit unterschiedlich zusammengesetzten Stichproben von Sexualstraftätern lagen die Raten für erneute schwerwiegende Delikte (Sexual- und/oder Gewaltdelikte) z. B. zwischen 14 % und 37 % (STADTLAND et al. 2006; HILL et al. 2008; EHER et al. 2008; ENDRASS et al. 2009). Mittels der Receiver-Operating-Curves (ROC) und den dazugehörigen AUCs lassen sich die untersuchten Risikoprognoseinstrumente hinsichtlich ihrer prädiktiven Validität vergleichen.

Eine AUC verbindet zwei Aspekte der Vorhersagevalidität eines Testverfahrens in einem Wert: Sensitivität und Spezifität (MOSSMAN 1994). Die Sensitivität (Richtig-Positiv-Rate) eines Risikoprognoseinstrumentes beschreibt den Anteil der rückfälligen Probanden, der korrekt als rückfällig klassifiziert wird. Dagegen beschreibt die Spezifität (Richtig-Negativ-Rate) den Anteil der nicht-rückfälligen Probanden, der korrekt als nicht-rückfällig klassifiziert wird. Zur Bestimmung der AUC trägt man für jeden Risikoscore die Richtig-Positiv-Rate auf der y-Achse mit der dazugehörigen Falsch-Positiv-Rate (1-Spezifität) auf der x-Achse ab. Die Fläche unterhalb dieser Kurve wird als AUC bezeichnet.

Zum Beispiel erzielte der STATIC-99 (HANSON & THORNTON 1999) in einer Auswertung des Münchener Prognoseprojektes (STADTLAND et al. 2006) eine AUC von .72 ($p < .001$), d. h. ein zufällig ausgewählter rückfällig gewordener Proband hat mit einer Wahrscheinlichkeit von 72 % einen höheren STATIC-99-Score als ein zufällig ausgewählter nicht-rückfällig gewordener Proband. Was sagt ein AUC-Wert über die individuelle Wahrscheinlichkeit einer erneuten Straftat aus? Die AUC alleine sagt nicht, um wie viel höher der Testscore ist, wie hoch die Treffer- oder Fehlerquote insgesamt liegt, und vor allem nicht, wie hoch die individuelle Wahrscheinlichkeit ist, bei einem bestimmten Testscore erneut eine Straftat zu begehen. Selbst RICE und HARRIS (1995), die die AUC in der Risikoprognoseforschung propagieren, bemerken kritisch, dass ein einziger statistischer Kennwert, bzw. die AUC, nichts über die Brauchbarkeit eines Risikoprognoseinstrumentes im Einzelfall aussagt. Der Standardfehler der AUC und der Kurvenverlauf sind nach RICE und HARRIS (1995) weitere Kriterien, die bei der Bewertung eines Prognoseinstrumentes berücksichtigt werden müssen. Eine statistisch hoch signifikante AUC muss darüber hinaus nicht zwangsläufig eine praktische Relevanz haben. JACOBSON und TRUAX (1991) weisen darauf hin, dass eine hoch signifikante Effektstärke relativ unabhängig von deren klinischer Relevanz ist. Die Frage nach dem praktischen Nutzen eines signifikanten Ergebnisses lässt sich statistisch nicht beantworten, sondern nur anhand allgemeingültiger Konventionen. In der klinischen chemischen Forschung wird zum Beispiel ein AUC-Wert von .60 als »ungenügend«, von .70 als »mangelhaft«, von .80 als »ausreichend«, von .90 als »hoch« und von .95 als »nahezu perfekt« bewertet (OBUCHOWSKI, LIEBER & WIANS 2004). Diese relativ konservative Konvention wurde weitgehend auch von SJÖSTEDT und GRANN (2002) angenommen, wobei die Autoren gleichzeitig darauf hinweisen, dass die Bedeutung der AUC in der Bewertung von Risikoprognoseinstrumenten überschätzt wird. In ihrem Übersichtsartikel zur Gefährlichkeitsprognostik sehen DOLAN und DOYLE (2000) bereits eine AUC von $>.75$ als hoch an. Welche

Konventionen für die forensische Prognoseforschung als geeignet erachtet werden können, muss an anderer Stelle diskutiert werden.

Zusammenfassend bietet die AUC zwar eine elegante statistische Lösung für den Vergleich der prädiktiven Validität von Risikoprognoseinstrumenten, die in Populationen mit unterschiedlichen Basisraten angewandt wurden. Die größte Bedeutung hat die AUC jedoch bei der Bestimmung des optimalen »Cut-Off-Scores«: Anhand des Kurvenverlaufs der ROC lässt sich der Risikoscore bestimmen, der die niedrigste Rate an Falsch-Positiven mit der höchstmöglichen Rate an Richtig-Positiven verbindet. Für den Praktiker hat die AUC jedoch kaum eine Bedeutung, da die Beantwortung der Frage nach der Wahrscheinlichkeit für einen höheren Risikoscore bei einem rückfälligen bzw. einem niedrigen Risikoscore bei einem nicht-rückfälligen Probanden nichts über die individuelle Gefährlichkeit eines Probanden aussagt.

Heterogene Ausgangsstichproben – Repräsentativ für den Einzelfall?

Eine weitere zentrale Problematik der derzeitigen empirischen Evaluation von Risikoprognoseinstrumenten ist die Heterogenität der Ausgangsstichproben (DOLAN & DOYLE 2000). Die Mehrheit der Validitätsstudien nutzt Stichproben, die sich aus Straftätern mit den verschiedensten Delikttypen, klinischen Störungsbildern und unterschiedlichsten Altersgruppen zusammensetzen. Häufig handelt es sich um sogenannte anfallende Stichproben, d. h. ein Risikoprognoseinstrument wurde zum Beispiel anhand von vorliegenden Akteninformationen ausgefüllt und retrospektiv genutzt. Neben der Heterogenität der Stichproben sind die je nach Untersuchung stark variierenden Prognosezeiträume (»time at risk«), die uneinheitlich gewählten Cut-Off-Scores (z. B. URBANIOK 2004; FREEDMAN 2001) und die unterschiedlichen Definitionen von Rückfälligkeit (z. B. SJÖSTEDT & GRANN 2002) sowie das Übergehen potenzieller Moderatorvariablen und protektiver Faktoren (z. B. ROGERS 2000) als problematisch zu bewerten.

Für die Gefährlichkeitsprognose des Einzelfalls werden jedoch spezifische statistische Bezugsnormen benötigt, d. h. die Normstichprobe muss, wie auch bei anderen psychometrischen Verfahren, in wesentlichen Kriterien dem jeweiligen Einzelfall entsprechen, um eine valide Prognose zu gewährleisten. Wesentliche Kriterien für eine Normierung von Risikoprognoseinstrumenten können neben der Art des Indexdelikts, einem definierten Prognosezeitraum und einer einheitlichen Kriteriumsvariable (z. B. erneute Delinquenz mit körperlicher Gewaltanwendung) auch Faktoren wie die aktuelle klinische Diagnose, das Alter und das bei Entlassung in die Freiheit zu erwartende soziale Setting sein. Für die Praxis bedeutet dies, dass die Bestimmung von Risikoscores für Lockerungsentscheidungen zumindest aus methodischer Sicht unzulässig sind, da bisher kein Risikoprognoseinstrument existiert, welches anhand einer repräsentativen Straftäterpopulation hinsichtlich der Kriteriumsvariable »Lockerung« (z. B. Ausgang im Gelände) normiert wurde.

In den formulierten Mindestanforderungen für die Erstellung von Prognosegutachten (BOETTICHER et al. 2006) wird darauf hingewiesen, dass Prognoseinstrumente sinnvolle Checklisten

sein, aber nicht alleine die Grundlage für die Prognose weiterer Straftaten bilden können. Darüber hinaus wird gefordert, dass die verwendeten Verfahren bereits aus ethischen Gründen methodische Mindestanforderungen erfüllen müssen: Standardisierung, wie Instruktionen zur Anwendung, zur Itemcodierung und zur Auswertung. Es müssen Erkenntnisse zur Reliabilität und zur Validität vorliegen. Zu diesen methodischen Mindestanforderungen ließe sich nach den im vorangegangenen Abschnitt diskutierten Problemen noch die zusätzliche Forderung nach einer Normierung stellen. Denn ohne repräsentative statistische Bezugsnormen lassen selbst reliable und standardisierte empirische Verfahren keine für den Einzelfall validen Prognosen zu. Des Weiteren würden einheitliche Normierungsstandards die Vergleichbarkeit von Risikoprognoseinstrumenten erhöhen.

Basisratenproblematik – Einflüsse der Ausgangswahrscheinlichkeiten auf individuelle Rückfallwahrscheinlichkeiten

Die sogenannte »base rate fallacy« (Basisratentäuschung) erlangte in den 1970er-Jahren insbesondere durch KAHNEMAN und TVERSKY (1973) in der empirischen Forschung Bedeutung. TVERSKY und KAHNEMAN (1982) postulieren, dass bei einer Vielzahl heuristischer Entscheidungen die a priori Basisraten eines Ereignisses unberücksichtigt bleiben, was insbesondere bei seltenen Ereignissen zu Fehlentscheidungen führen kann. Für die forensische Risikoprognoseforschung wurde diese Problematik kürzlich von DONALDSON und WOLLERT (2008) diskutiert und ein auf dem Bayes Theorem basierendes statistisches Vorgehen dargestellt. Auch VOLCKART (1999), DAHLE (2006 b) und NEDOPIL (2005) weisen auf die Bedeutung der Basisraten bei der Erstellung individueller Risikoprognosen hin.

Folgt man dem Satz von Bayes (z. B. LYNCH 2007), so muss man neben der Falsch- ($p(\text{test+}|\text{rück-})$) und Richtig-Positiv-Rate ($p(\text{test+}|\text{rück+})$) eines Risikoprognoseinstrumentes auch die Basisrate ($p(\text{rück+})$) der Delikte, die vorhergesagt werden sollen, berücksichtigen. Die Frage lautet somit: Wie hoch ist die Wahrscheinlichkeit, dass ein als rückfällig klassifizierter Proband tatsächlich rückfällig wird ($p(\text{rück+}|\text{test+})$)?

$$p(\text{rück+}|\text{test+}) = \frac{p(\text{test+}|\text{rück+}) \cdot p(\text{rück+})}{p(\text{test+}|\text{rück+}) \cdot p(\text{rück+}) + p(\text{test+}|\text{rück-}) \cdot p(\text{rück-})}$$

Was bedeutet dies für die bei Risikoprognoseinstrumenten angegebenen Rückfallwahrscheinlichkeiten und die individuelle Gefährlichkeitsbeurteilung?

Zum Beispiel wurde in einer Untersuchung von SJÖSTED und LÄNGSTRÖM (2001) die prädiktive Validität des STATIC-99 (HANSON & THORNTON 1999) an einer Stichprobe von 1 368 Straftätern untersucht. Die durchschnittliche Katamnesedauer betrug 3,7 ($SD = 1,4$) Jahre. Der mithilfe der ROC-Analyse und der AUC bestimmte optimale Cut-Off-Score war ≥ 4 Punkte ($n = 190$; 14%). Für erneute Sexualdelikte betragen die Spezifität ($p(\text{test-}|\text{rück-})$) 88% und die Sensitivität ($p(\text{test+}|\text{rück+})$) 49%. Die Basisrate ($p(\text{rück+})$) für eine erneute Sexualstraftat lag in der Gesamtstichprobe bei 4%. Würde man diese Basisrate als repräsentativ ansehen und sie im Sinne des Bayes Theorems zur Bestimmung der Rückfallwahrschein-

70

lichkeit berücksichtigen, dann läge die Wahrscheinlichkeit für ein erneutes Sexualdelikt eines Probanden mit einem STATIC-99-Score ≥ 4 bei 15 %.

$$p(\text{rück+} | \text{test+}) = \frac{(.49) * (.04)}{(.49) * (.04) + (.12) * (.96)} = .15$$

Anders formuliert würde man seine prognostische Entscheidung ausschließlich von einem STATIC-99-Score abhängig machen, dann würde man 85 % der Probanden mit einem STATIC-99-Score ≥ 4 eine hohe Rückfallwahrscheinlichkeit unterstellen, obwohl sie dem Bayes Theorem folgend keinen Rückfall begehen werden. Eine Rückfallwahrscheinlichkeit von 15 % ist wiederum etwa viermal so hoch wie die Basisrate innerhalb der Gesamtstichprobe (4 %). Die Wahrscheinlichkeit ($p(\text{rück-} | \text{test-})$), dass ein Proband mit einem STATIC-99-Score < 4 tatsächlich keinen Rückfall begeht, liegt dagegen bei 98 %. Die Wahrscheinlichkeit, dass ein als nicht-rückfällig klassifizierter Proband einen Rückfall begeht, beträgt somit 2 % und ist halb so hoch wie die Basisrate (4 %) in der Gesamtstichprobe. Dieses Rechenbeispiel verdeutlicht den Einfluss der Basisraten auf die prädiktive Validität von Risikoprognoseinstrumenten und dies trotz der geringen Rate von Falsch-Positiven ($p(\text{test+} | \text{rück-}) = 12\%$).

Zur Verdeutlichung wird das oben genannte Beispiel in Abbildung 1 in Form eines Wahrscheinlichkeitsbaums dargestellt. $P(\text{rück+} \cap \text{test+})$ gibt hierbei die Wahrscheinlichkeit an, dass ein Proband rückfällig wird *und* einen hohen STATIC-99-Score hat. $P(\text{rück+} \cap \text{test-})$ ist die Wahrscheinlichkeit, dass ein Proband rückfällig wird *und* einen niedrigen STATIC-99-Score hat. Entsprechend beschreibt $p(\text{rück-} \cap \text{test+})$ die Wahrscheinlichkeit, dass ein Proband nicht rückfällig wird *und* einen ho-

hen STATIC-99-Score hat und $p(\text{rück-} \cap \text{test-})$ die Wahrscheinlichkeit, dass ein Proband nicht rückfällig wird *und* einen niedrigen STATIC-99-Score hat. Aus dem Bayes Theorem ergibt sich folgendes: Die Wahrscheinlichkeit, dass ein Proband rückfällig wird und einen hohen STATIC-99-Score hat, muss durch die Wahrscheinlichkeit eines hohen STATIC-99-Scores ($p(\text{test+}) = p(\text{rück+} \cap \text{test+}) + p(\text{rück-} \cap \text{test+})$) geteilt werden, um die Wahrscheinlichkeit zu erhalten, dass ein Proband mit einem hohen STATIC-99-Score tatsächlich rückfällig wird (s. Formel »Relevanz«, Abb. 1). Man stellt nun fest, dass sich aufgrund der geringen Basisrate (4 %) Probanden mit »hohem« STATIC-99-Score fast sechsmal häufiger in der nicht rückfälligen als in der rückfälligen Gruppe finden (11 % vs. 2 %). Die AUC betrug in der dargestellten Studie von SJÖSTED und LÅNGSTRÖM (2001) .76 (CI-95 % .69 – .83) und war mit $p < .01$ hoch signifikant. Dieses Beispiel verdeutlicht, dass die AUC als alleiniger Validitätskoeffizient oder gar als Goldstandard für die Risikoprognoseforschung unbrauchbar ist.

Von welcher a priori Basisrate von Rückfälligkeit im Einzelfall auszugehen ist, lässt sich nur schwer ermitteln. Basisraten unberücksichtigt zu lassen oder von einer Bernoulli-Verteilung (Null-Eins-Verteilung) auszugehen, in der die Ereignisse »Rückfall« und »kein Rückfall« gleichwahrscheinlich sind, kann jedoch zu deutlichen Fehleinschätzungen (i. d. R. einer Überschätzung) der prädiktiven Validität von aktuarischen Risikoprognoseinstrumenten führen. Nach DAHLE (2006b) sollten Basisraten so gewählt werden, dass sie zumindest im Hinblick auf das Geschlecht, die Altersgruppe und die Art und Schwere des Anlassdelikts den Eigenschaften des zu beurteilenden Probanden entsprechen. In einer Übersichtsarbeit von GROSS und NEDOPIL (2005) werden für unterschiedliche Delikttypen gängige Basisraten aus der Literatur zusammengefasst.

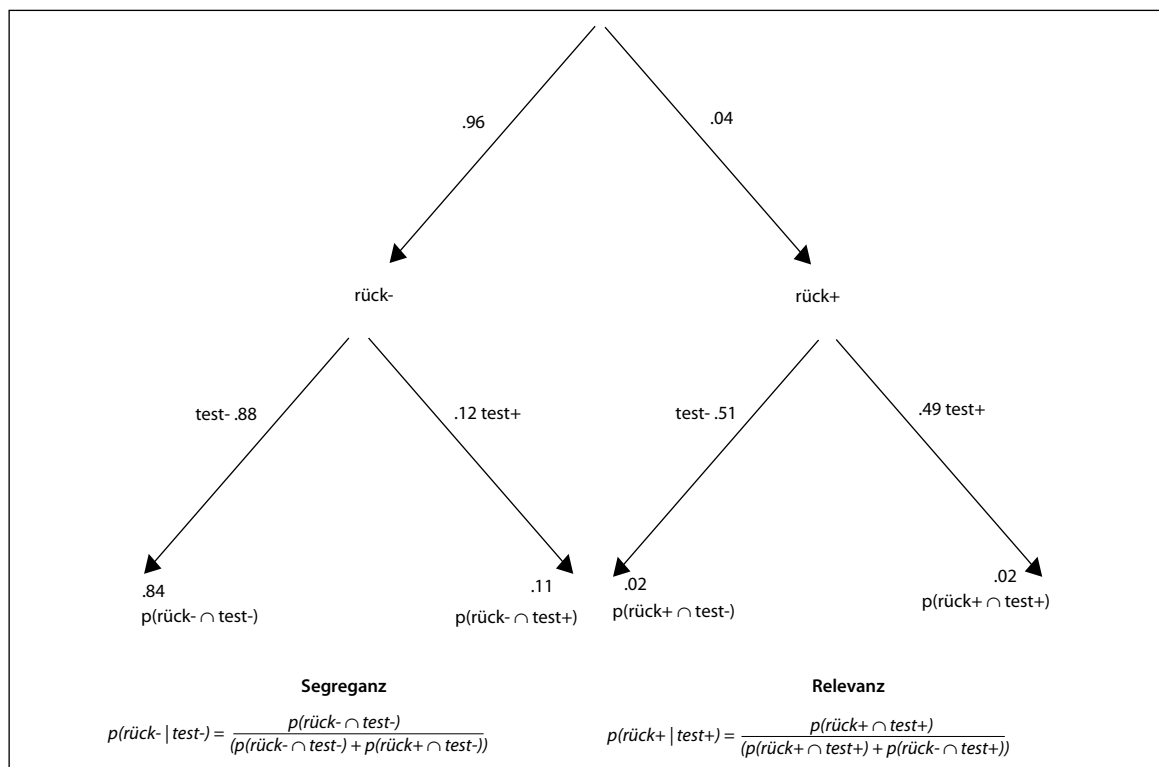


Abb. 1: Wahrscheinlichkeitsbaum nach dem Bayes Theorem

Bemerkung: rück+ = rückfällig; rück- = nicht rückfällig; test+ = hoher STATIC-99-Score ≥ 4 ; test- = niedriger STATIC-99-Score < 4 . Die Daten im dargestellten Beispiel stammen aus SJÖSTED & LÅNGSTRÖM (2001).

Neben den Basisraten spielt auch die Selektionsrate eine wesentliche Rolle bei der Beurteilung der Validität von aktuari-schen Risikoprognoseinstrumenten (DAHLE 2006b). Unter der Selektionsrate versteht man den Anteil an Probanden, der aufgrund eines hohen Risikoscores nicht in die Freiheit entlassen werden kann und somit keine bzw. nur eine bedingte Möglichkeit hat, Delikte zu begehen. Diese Probandengruppe wäre somit zu der Gruppe der Richtig-Positiven zu zählen. Auch wenn der situative Rahmen Delikte erschwert, schließen eine Inhaftierung oder eine Unterbringung im Maßregelvollzug erneute schwere Straftaten nicht vollkommen aus. Auf das zuvor dargestellte Beispiel trifft die Problematik der Selektionsrate jedoch nicht zu, da es sich in der Untersuchung von SJÖSTED und LÅNGSTRÖM (2001) um eine Population von Straftätern handelt, die nach Verbüßung ihrer Haftstrafe in Freiheit entlassen wurden.

Für die Validierung von Risikoprognoseinstrumenten bedeutet dies, dass neben der AUC, der Sensitivität und der Spezifität eines Risikoprognoseinstrumentes auch die klassischen Testgütekriterien Segreganz ($p(\text{rück-}|\text{test-})$) und ganz besonders Relevanz ($p(\text{rück+}|\text{test+})$) mit den verwendeten Basisraten angegeben werden sollten, da diese für den Praktiker einen höheren Informationsgehalt haben (HART, WEBSTER & MENZIES 1993). Insbesondere hohe Risikoscores sollten in der Praxis einer gründlichen klinischen Prüfung unterzogen werden, in der es gilt, die Bedeutung der erhobenen Items für den jeweiligen Probanden zu diskutieren.

Risikoprognoseinstrumente – valide Grundlage für Einzelfallentscheidungen?

Der praktische Nutzen von Risikoprognoseinstrumenten besteht vorwiegend darin, dass sie Items bzw. innerpsychische, klinische und/oder soziodemografische Faktoren beinhalten, die mit Rückfälligkeit assoziiert sind. Inwieweit diese Faktoren auch im jeweiligen Einzelfall von Bedeutung sind, kann nur anhand eines individuellen strukturierten klinischen (»structured clinical judgement«) Vorgehens beurteilt werden (z. B. DAHLE 1997; DAHLE 2006b; DIETIKER, DITTMANN & GRAF 2007). Der Ansatz der individuellen strukturierten klinischen Beurteilung versucht die Lücke zwischen rein aktuari-schen Risikoprognoseinstrumenten und der eher intuitiven klinischen Praxis zu schließen (z. B. DOLAN & DOYLE 2002). Das primäre Ziel einer strukturierten klinischen Beurteilung des Einzelfalls ist nicht, erneute Straftaten vorherzusagen, sondern individuelle Risikomerkmale sowie protektive Faktoren zu identifizieren und daraus entsprechende Interventionsmöglichkeiten abzuleiten. Im Prozess der Identifizierung von individuellen Risikomerkmale können Risikoprognoseinstrumente eine sinnvolle Rolle spielen. Individuelle Interventionsmöglichkeiten lassen sich anhand von Risikoscores jedoch nicht ableiten.

Aufgrund der beschriebenen methodischen Probleme bei der Konstruktion von Risikoprognoseinstrumenten ist das Aufsummieren von Risikoscores daher kein geeignetes Vorgehen für die Bestimmung einer individuellen Rückfallwahrscheinlichkeit. In der aktuellen Risikoprognoseforschung wird davon ausgegangen, dass es sich bei »Rückfallrisiko« um ein lineares Konstrukt handelt: Je mehr ungünstige Faktoren ein Proband erfüllt, desto höher ist seine Rückfallwahrscheinlichkeit oder Gefährlichkeit zu beurteilen. MÜLLER-ISBERNER, JÖCKEL und

CABEZA (1998) weisen in ihrem Handbuch zum HCR-20 richtigerweise darauf hin, dass bereits das Vorliegen eines einzelnen Items bei einem Probanden für eine individuelle Gefährlichkeit sprechen kann und dass es sich beim HCR-20 um eine Anleitung zur Untersuchung von Probanden mit gewalttätigen Verhalten in der Vorgeschichte handelt, bei denen der Verdacht auf eine psychische Erkrankung oder eine Persönlichkeitsstörung besteht. Dabei ist auch die Frage, ob sich das empirische Konstrukt »Rückfallrisiko« tatsächlich am besten durch lineare multivariate Modelle statistisch beschreiben lässt, bis dato noch ungeklärt.

Andere statistische Modelle wie z. B. die »Classification and Regression Tree Analysis (CART)« (STEADMAN 2000) scheinen gewisse Vorteile gegenüber linearen Modellen zu haben. Ein Vorteil liegt darin, dass nicht von der Unabhängigkeit der einzelnen Risikofaktoren (Items des Instrumentes) ausgegangen wird, sodass Probanden anhand unterschiedlichster Variablenkombinationen verschiedenen Risikosubgruppen zugeordnet werden können. Die CART-Analyse fokussiert auf Interaktionen anstatt auf Haupteffekte einzelner Risikoprädiktoren (MONAHAN et al. 2005). Die Prädiktorvariablen können ein beliebiges Skalenniveau aufweisen und die Kriteriumsvariable muss keine dichotome Ausprägung haben, sondern kann mehrstufig sein (z. B. »keine Straftat«, »Delikte ohne Gewalt« und »Gewaltdelikte«). Da es sich um ein non-parametrisches Verfahren handelt, werden auch geringere Forderungen bezüglich der Verteilung gestellt. Darüber hinaus können zur Bestimmung der »Classification-Trees« sowohl die Basisraten berücksichtigt werden als auch die »Kosten« einer Fehlklassifikation. Einen Probanden, der ein Gewaltdelikt begehen wird, einer niedrigen Risikokategorie zuzuordnen, wäre eine schwerwiegendere Fehlentscheidung, als ihm ein mittleres Risiko zuzuschreiben. Den größten Vorteil bietet die auch für Laien leichte Interpretierbarkeit der »Classification-Trees«, sodass sich die Ergebnisse einfacher in der klinischen Praxis umsetzen lassen. STEADMAN et al. (2000) führten einen direkten Vergleich zwischen Risikoprognoseinstrumenten basierend auf multivariaten linearen Modellen und auf der CART-Analyse durch. In Bezug auf die AUCs waren die Unterschiede zwischen den beiden statistischen Methoden marginal. Der Anteil von Probanden, deren vorhergesagte Rückfallwahrscheinlichkeit in etwa der Basisrate entsprach, lag mit 23,5 % für die CART-Analyse jedoch deutlich niedriger als bei den auf Haupteffekten basierenden Verfahren (42,9 % bis 49,2 %). DAHLE (2006a) spricht hier von der Gruppe der sogenannten »unpredictables« (Nicht-Vorhersagbare Probanden), d. h. für etwa die Hälfte aller Probanden bieten Risikoprognoseinstrumente, die auf linearen Kombinationen basieren, keinen zusätzlichen Informationsgewinn.

Was gilt es bei der Auswahl eines geeigneten Risikoprognoseinstrumentes für die Einzelfallentscheidung zu beachten? Der Praktiker sollte zunächst folgende Fragen klären:

- || Ist der zu beurteilende Proband in wesentlichen Kriterien (z. B. Anlassdelikt, psychische Erkrankung, Alter) vergleichbar mit der Normierungsstichprobe des angewendeten Risikoprognoseinstrumentes?
- || Ist das für den Probanden zu erwartende Setting vergleichbar mit dem der Normierungsstichprobe des angewendeten Risikoprognoseinstrumentes?
- || Ist der Zeitraum (»time at risk«), über den eine Risikoprognose erstellt werden soll, vergleichbar mit dem Zeitraum

der Normierungsstichprobe des angewendeten Risikoprognoseinstrumentes?

Falls keine oder eine geringe Vergleichbarkeit zwischen der Normierungsstichprobe des angewendeten Risikoprognoseinstrumentes und dem zu beurteilenden Einzelfall besteht, ist die Bestimmung eines individuellen Risikoscores aus methodischer Sicht nicht zu rechtfertigen. In diesem Fall können Risikoprognoseinstrumente, wie kürzlich von BOETTICHER et al. (2009) formuliert, lediglich als Checklisten dienen, da eine unkritische Übernahme gruppenstatistischer Erkenntnisse keine empirisch begründete Wahrscheinlichkeitsaussage für den Einzelfall zulässt. Neben der Vergleichbarkeit müssen auch die Basisraten für die vorherzusagenden Delikte berücksichtigt werden. Je geringer die Basisrate eines bestimmten Delikts ist, desto stärker fällt die Falsch-Positiv-Rate, also der Anteil an Probanden, der fälschlicherweise als rückfällig klassifiziert wird, ins Gewicht. Ein hoher Anteil an Falsch-Positiven ist ein Problem, welches in einer Vielzahl von Risikoprognosestudien auftritt (z. B. NEDOPIL & STADTLAND 2007), sodass die Relevanz (s. Tab. 1) von Risikoprognoseinstrumenten meist deutlich unter 50 % liegt, was heißt, dass die Mehrheit der als hoch rückfallgefährdet klassifizierten Probanden tatsächlich keinen Rückfall begeht. Im anglo-amerikanischen Raum hat sich nach ACKERMAN (1999) in der forensischen Praxis das 51. Perzentil als Standard für eine »hinreichend hohe Wahrscheinlichkeit« etabliert: wenn ein Ereignis in mehr als 50 % der Fälle zu erwarten ist, geht man von einer für den Einzelfall hinreichend hohen Wahrscheinlichkeit aus. Aktuarische Risikoprognoseinstrumente erfüllen diesen »Standard« für die Vorhersage von Straftaten in aller Regel nicht, da deren Relevanz meist unterhalb der 50%-Wahrscheinlichkeit liegt. Dagegen erweisen sich Risikoprognoseinstrumente als recht valide in Bezug auf ihre Segreganz (s. Tab. 1), d. h. ein als niedrig rückfallgefährdet klassifizierter Proband begeht mit einer hohen Wahrscheinlichkeit kein erneutes Delikt. Die Segreganz von aktuarischen Risikoprognoseinstrumenten überschreitet die 50%-Wahrscheinlichkeit meist deutlich. Das heißt für die Praxis: die verbreiteten Prognoseinstrumente können Nicht-Rückfälligkeit sicherer vorhersagen als Rückfälligkeit.

Bei der Verwendung aktuarischer Risikoprognoseinstrumente in forensisch-psychiatrischen Settings bleibt zusätzlich zu bedenken, dass die große Mehrheit der dort untergebrachten Patienten eine ungünstige psychosoziale Herkunft, eine Geschichte von Drogen- oder Alkoholmissbrauch und eine prekäre berufliche Entwicklung hat. Hier können aktuarische Risikoprognoseinstrumente wenig differenzieren, da die Varianz innerhalb dieser spezifischen Straftäterpopulation hinsichtlich historischer Items gering ist und so eine Aussage über die Wahrscheinlichkeit erneuter Straftaten in der Regel immer negativ ausfallen wird.

Ab welcher empirisch bestimmten Rückfallwahrscheinlichkeit ein Individuum als »hoch«, »moderat« oder »niedrig« gefährlich anzusehen ist, wird dabei immer eine politische, juristische und ethische Frage bleiben, die sich auf dem statistischen Weg nicht beantworten lässt. Hierzu gilt es Risiken, wie den potenziellen Schaden, den Opfer durch ein erneutes schweres Delikt erleiden, den Folgen eines zu Unrecht untergebrachten Straftäters gegenüberzustellen und zu bewerten. Auch in Zukunft wird die forensische Risikoprognoseforschung zu dieser Fragestellung keinen Beitrag leisten können.

Literatur

- ACKERMAN MJ (1999) *Essentials of Forensic Psychological Assessment*. New York: John Wiley & Sons
- ANDREWS DA, BONTA J (2003) *The Psychology of Criminal Conduct*. Cincinnati: Anderson Publishing Co.
- BOETTICHER A, KRÖBER HL, MÜLLER-ISBERNER R, BÖHM KM, MÜLLER-METZ R, WOLF T (2006) Mindestanforderungen für Prognosegutachten. *Neue Zeitschrift für Strafrecht* 26: 537–592
- BOETTICHER A, DITTMANN V, NEDOPIL N, NOWARA S, WOLF T (2009) Zum richtigen Umgang mit Prognoseinstrumenten durch psychiatrische und psychologische Sachverständige und Gerichte. *Neue Zeitschrift für Strafrecht* 29: 478–481
- COOKE DJ, MICHIE C (2009) Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, 33: 541–556
- DAHLE KP (1997) Kriminalprognosen im Strafrecht: Psychologische Aspekte individueller Verhaltensvorhersagen. In: STELLER M, VOLBERT R (Hg.) *Psychologie im Strafverfahren*. Bern: Huber, 119–140
- DAHLE KP (2006 a) Strength and limitations of actuarial prediction of criminal reoffence in a German prison sample: A comparative study of LSI-R, HCR-20 and PCL-R. *International Journal of Law and Psychiatry* 29: 431–442
- DAHLE KP (2006 b) Grundlagen und Methoden der Kriminalprognose. In: KRÖBER HL, DÖLLING D, LEYGRAF N, SASS H (Hg.): *Handbuch der Forensischen Psychiatrie – Psychiatrische Kriminalprognose und Kriminaltherapie* (Bd. 3). Darmstadt: Steinkopff, 1–67
- DAHLE KP, SCHNEIDER V, ZIETHEN F (2006) Standardisierte Instrumente zur Kriminalprognose. *Forensische Psychiatrie, Psychologie, Kriminologie* 1: 15–26
- DIETIKER J, DITTMANN V, GRAF M (2007) Gutachterliche Risikoeinschätzung bei Sexualstraftätern – Anwendbarkeit von PCL-SV, HCR-20+3 und SVR-20. *Nervenarzt* 78: 53–61
- DOLAN M, DOYLE M (2000) Violence risk prediction: Clinical and actuarial measures and the role of the Psychopathy Checklist. *British Journal of Psychiatry* 177: 303–311
- DOLAN M, DOYLE M (2002) Violence risk assessment: Combining actuarial and clinical information to structure clinical judgements for the formulation and management of risk. *Journal of Psychiatric and Mental Health Nursing* 9: 649–657
- DONALDSON T, WOLLERT R (2008) A mathematical proof and example that Bayes's theorem is fundamental to actuarial estimates of sexual recidivism risk. *Sexual Abuse: A Journal of Research and Treatment* 20: 206–217
- EHER R, RETTENBERGER M, SCHILLING F, PFÄFFLIN F (2008) Failure of STATIC-99 and SORAG to predict relevant reoffense categories in relevant offender subtypes: A prospective study. *Sexual Offender Treatment* 3: 1–14
- ENDRASS J, URBANIOK F, HELD L, VETTER S, ROSSEGGER A (2009) Accuracy of the Static-99 in predicting recidivism in Switzerland. *International Journal of Offender Therapy and Comparative Criminology* 53: 482–490
- FREEDMAN D (2001) False prediction of future dangerousness: Error rates and Psychopathy Checklist-Revised. *The Journal of the American Academy of Psychiatry and the Law* 29: 89–95
- GROSS G, NEDOPIL N (2005) Basisraten für kriminelle Rückfälle – Ergebnisse einer Literaturübersicht. In: NEDOPIL N (Hg.) *Prognosen in der forensischen Psychiatrie – ein Handbuch für die Praxis*. Lengerich: Pabst Science Publisher, 65–98

- HANSON RK, THORNTON D (1999) Static 99: Improving actuarial risk assessments for sex offenders (User Report No. 1999-02). Ottawa: Department of the Solicitor General of Canada
- HART SD, MICHIE C, COOKE DJ (2007) Precision of actuarial risk assessment instruments: Evaluating the margins of error of group v. individual predictions of violence. *British Journal of Psychiatry* 190: 60–65
- HART SD, WEBSTER CD, MENZIES RJ (1993) A note on portraying the accuracy of violence predictions. *Law and Human Behavior* 17: 695–700
- HILL A, HABERMANN N, KLUSMANN D, BERNER W, BRIKEN P (2008) Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy and Comparative Criminology* 52: 5–20
- JACOBSON NS, TRUAX P (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Research* 59: 12–19
- KAHNEMAN D, TVERSKY A (1973) On the psychology of prediction. *Psychological Review* 80: 237–351
- LYNCH SM (2007) *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer
- MELOY JR (1992) Discussion of »On the predictability of violent behavior: considerations and guidelines.« *Journal of Forensic Sciences* 37: 949–950
- MONAHAN J, STEADMAN HJ, ROBBINS PC, APPELBAUM P, BANKS S, GRISSO T, HEILBRUN K, MULVEY EP, ROTH L, SILVER E (2005) An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services* 56: 810–815
- MOSSMAN D (1994) Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology* 62: 783–792
- MÜLLER-ISBERNER R, JÖCKEL D, CABEZA SG (1998) Die Vorhersage von Gewalttaten mit dem HCR-20 (The Prediction of Violence with the HCR-20 Scheme). Haina, GER: Institut für Forensische Psychiatrie
- NEDOPIL N (2007) *Forensische Psychiatrie: Klinik, Begutachtung und Behandlung zwischen Psychiatrie und Recht*. Stuttgart: Georg Thieme Verlag KG, 188–191
- NEDOPIL N, STADTLAND C (2007) Das Problem der falsch Positiven: Haben wir unsere prognostische Kompetenz seit 1966 verbessert? In: LÖSEL F, BENDER D, JEHLE JM (Hg.) *Kriminologie und wissenschaftsbasierte Kriminalpolitik: Entwicklungs- und Evaluationsforschung*. Mönchengladbach: Forum Verlag Godesberg GmbH, 541–550
- OBUCHOWSKI NA, LIEBER ML, WIANS FH (2004) ROC curves in clinical chemistry: Uses, misuses and possible solutions. *Clinical Chemistry* 50: 1118–1125
- QUINSEY VL, HARRIS G, RICE M, CORMIER CA (2003) *Violent Offenders: Appraising and Managing Risk*. Washington, DC: American Psychological Association
- RICE ME, HARRIS GT (1995) Violent recidivism: assessing predictive validity. *Journal of Consulting and Clinical Psychology* 63: 737–748
- ROGERS R (2000) The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior* 24: 595–605
- SJÖSTEDT G, GRANN M (2002) Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health* 1: 179–183
- SJÖSTED G, LÅNGSTRÖM N (2001) Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior* 25: 629–645
- STADTLAND C, HOLLWEG M, KLEINDIENST N, DIETL J, REICH U, NEDOPIL N (2006) Rückfallprognose bei Sexualstraftätern – Vergleich der prädiktiven Validität von Prognoseinstrumenten. *Nervenarzt* 77: 587–595
- STEADMAN HJ, SILLVER E, MONAHAN J, APPELBAUM PS, ROBBINS PC, MULVEY EP, GRISSO T, ROTH LH, BANKS S (2000) A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior* 24: 83–100
- TVERSKY A, KAHNEMAN D (1982) Evidential Impact of Base Rates. In: KAHNEMAN D, SLOVIC P, TVERSKY A (Hg.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press
- URBANIOK F (2004) Validität von Risikokalkulationen bei Straftätern – Kritik an einer methodischen Grundannahme und zukünftige Perspektiven. *Fortschritte der Neurologie, Psychiatrie* 72: 260–269
- VOLCKART B (1999) Zur Bedeutung des hermeneutischen Verstehens in der Kriminalprognose. *Recht & Psychiatrie* 17: 58–64
- WALSH T, WALSH Z (2006) The evidentiary introduction of Psychopathy Checklist-Revised assessed psychopathy in U.S. courts: Extent and appropriateness. *Law and Human Behavior* 30: 493–507
- WEBSTER CD, DOUGLAS KS, EAVES D, HART SD (1997) *HCR-20: Assessing Risk of Violence (Version 2)*. Vancouver, CA: Mental Health Law and Policy, Institute Simon Fraser University

Anschrift des Verfassers

Universität Duisburg-Essen
 Institut für Forensische Psychiatrie
 LVR Klinikum Essen
 Virchowstr. 174
 45147 Essen
 andrej.koenig@uni-due.de
