

Medizinische Universitäts-Poliklinik, Kantonsspital Basel¹, Praxis für Allgemeinmedizin und Institut für klinische Epidemiologie, Einsiedeln², Medizinische Poliklinik, Departement Innere Medizin, Universitätsspital Zürich³

¹H.C. Bucher, ²J.G. Schmidt, ³J. Steurer

Kritische Beurteilung einer Arbeit zu einem diagnostischen Test

Critical Appraisal of a Publication on a Diagnostic Test

Klinisches Szenario

Sie betreuen einen 55jährigen Schichtarbeiter, der unter einer chronisch obstruktiven Lungenkrankheit (COPD) leidet. Der Patient ist übergewichtig (Body mass index [BMI] 28,0 kg/m²) und Raucher. Er leidet unter einer mässig befriedigend eingestellten Hypertonie, und Sie vermuten einen Äthylabusus. Der Patient hat Ihnen gegenüber auch schon Potenzprobleme geäussert. Während Ihrer Ferienabwesenheit sucht der Patient wegen einer Exazerbation seiner COPD die Praxis eines spezialisierten Kollegen auf. Der Pneumologe erfährt im Gespräch von Einschlafneigungen am Tag und veranlasst eine ambulante nächtliche Pulsoxymetrie, welche pathologisch ausfällt. Aufgrund der typischen Symptomenkonstellation, die Lebenspartnerin habe sich über das Schnarchen des Patienten schon beklagt, schlägt er eine Abklärung mittels Polysomnographie vor. Dem Bericht des Kollegen entnehmen Sie, dass die ambulante Pulsoxymetrie einen Entsättigungsindex von 12 pro Stunde (Normalwert <10 pro Stunde) zeigte. Eine O₂-Entsättigung von >4% wird in der Befundbeilage als pathologisch definiert. In einer Rückfrage stuft Ihr Kollege die Chance eines Schlafapnoe-Syndroms bei diesem Patienten auf mindestens 50% ein.

Die Evidenz suchen

Ihr Plan ist es, Information über den Stellenwert der ambulanten Pulsoxymetrie zur Identifikation von Patienten

mit Verdacht auf Schlafapnoe-Syndrom zu suchen und das Problem in Ihrem Qualitätszirkel vorzustellen. Sie sind mit Medline vertraut und suchen mit dem Suchsystem Thesaurus den indextierten Suchbegriff (Medical Subject Headings, MeSH) «sleep apnea syndromes». Als zweiten MeSH-Begriff geben Sie «predictive value of tests» ein. Die Kombination dieser beiden Suchbegriffe ergibt für die Jahre 1989 bis 1997 insgesamt 32 Arbeiten. Bei der Durchsicht finden Sie fünf Arbeiten, die von potentiell Interesse sind. Zwei Arbeiten (1, 2) evaluieren die Pulsoxymetrie simultan im Rahmen der Polysomnographie, drei Arbeiten (3, 4, 5) evaluieren die ambulante Pulsoxymetrie im Vergleich mit der Polysomnographie. Sie beschaffen sich eine Kopie von Re-

ferenz 5, da diese den diagnostischen Stellenwert der ambulanten Pulsoxymetrie in Zusammenhang mit klinischen Beschwerden speziell evaluiert. Im folgenden soll gezeigt werden, wie eine Studie zu einem diagnostischen Test kritisch evaluiert werden kann (6, 7). Tabelle 1 listet die Kriterien auf, anhand derer vorgegangen werden soll.

1. Sind die Ergebnisse der Studie valide?

– Gibt es einen unabhängigen, verblindeten Vergleich mit einem Standardtest («gold standard»)?

Um sich zu versichern, ob der Test und die Testergebnisse die «wahren Verhältnisse» widerspiegeln, muss der/die Le-

Tab. 1. Kriterien zur Beurteilung der Methodik einer Studie zu einem diagnostischen Test (Referenz 6, 7)

1. Sind die Ergebnisse der Studie valide?

- Gibt es einen unabhängigen, verblindeten Vergleich mit einem Standardtest («gold standard»)?
- Schloss die Studie ein genügend breites Spektrum von Patienten ein, an welchen der Test in der klinischen Praxis angewendet wird?
- Beeinflussten die Ergebnisse des zu evaluierenden Tests die Entscheidung, ob der Standardtest angewendet wurde?
- Wurde die Testmethodik genügend detailliert beschrieben, um die Replikation des Tests zu ermöglichen?

2. Was sind die Studienergebnisse?

- Werden «Likelihood ratios» der Testergebnisse bzw. Sensitivität und Spezifität angegeben oder die Daten zu deren Berechnung aufgelistet?

3. Sind die Ergebnisse für die Behandlung meiner Patienten nützlich?

- Ist die Reproduzierbarkeit und die Interpretation der Testergebnisse in meinem klinischen Umfeld gegeben?
- Können die Ergebnisse an meinem Patienten angewendet werden?
- Werden die Testergebnisse das Patientenmanagement beeinflussen?

Korrespondenzadresse: Dr. H.C. Bucher, Medizinische Universitäts-Poliklinik, Kantonsspital Basel, Petersgraben 4, 4031 Basel

ser/in überprüfen, inwiefern ein zu evaluierender Test mit einem angemessenen Referenztest oder -standard verglichen worden ist. Der Referenztest muss ein Standard sein («gold standard»), d.h. der Test mit der höchsten ausgewiesenen Zuverlässigkeit, wie z.B. Biopsie, Autopsie, oder noch besser den für den Patienten entschiedenen Langzeit-Follow-up. Wurde ein Referenztest gewählt, der unbefriedigend ist, dann sind die Studienergebnisse kaum brauchbar. In einer Studie wurden beispielsweise D-Dimere als Screeningtest zur Diagnose von tiefen Venenthrombosen des Unterschenkels mit der Duplexsonographie verglichen (8). Die Verlässlichkeit der Duplexsonographie ist jedoch im Vergleich zur Phlebographie bei der Diagnose der Unterschenkelthrombose geringer, somit ist die Aussagekraft dieser Studie beschränkt.

Falls der Referenztest sinnvoll erscheint, muss weiter überprüft werden, ob die Ergebnisse des Tests und der Referenztest unabhängig evaluiert wurden. D.h. Personen, welche z.B. die Ergebnisse des Tests beurteilen, müssen gegenüber den Ergebnissen des «gold standard» (und umgekehrt) verblindet sein. Dass dies wichtig ist, wissen wir aus der klinischen Erfahrung: Nachdem wir einen vergrößerten mediastinalen Lymphknoten in der Computertomographie gesehen haben, fällt uns die vorher übersehene Verschattung auf der Thoraxübersichtsaufnahme auf. Je grösser die Wahrscheinlichkeit, dass die Interpretation des Tests durch Kenntnis des Ergebnisses der Referenzuntersuchung beeinflusst werden kann, um so wichtiger ist die Verblindung.

Die Autoren haben die ambulante nächtliche Pulsoxymetrie mit der Polysomnographie verglichen. Nach welchen Kriterien ein obstruktives Schlafapnoe-Syndrom in der Polysomnographie definiert ist, wird in der Fachliteratur kontrovers diskutiert. Gewisse Autoren bezeichnen einen Apnoe-Hypnoe-Index in der Polysomnographie von >15 pro Stunde als klinisch relevant, da erfahrungsgemäss meistens nur Patienten mit einem solchen Index durch eine nasale Überdruckbeatmungstherapie mit einem C-PAP-Gerät (continuous positive airflow pressure) profitieren und eine Verbesserung ihrer Einschlafneigungen erfahren. In der vorliegenden Studie wurde ein Apnoe-Hypnoe-Index von >10 pro Stunde in der Polysomnographie als

«pathologisch» definiert. Noch entscheidender ist die Tatsache, dass die klinische Bedeutung des Schlafapnoe-Syndroms ungenügend evaluiert ist und es unklar bleibt, ob es sich dabei um eine echte Krankheit und nicht um eine Labordiagnose handelt. Studien, welche einen Zusammenhang zwischen dem Schlafapnoe-Syndrom als Risikofaktor für Hypertonie, Herz- und Hirninfarkt, Gesamtsterblichkeit oder erhöhter Inzidenz von Verkehrsunfällen untersuchen, sind methodisch ungenügend und lassen keine schlüssige Bedeutung über die klinische Relevanz dieses Syndroms zu. Dem sogenannten Schlafapnoe-Syndrom kann hingegen eine Bedeutung bei gewissen Formen von Einschlafneigungen zukommen. Für die klinische Beurteilung dieser Störung wäre dann die mit Schlafapnoen verbundene Einschlafneigung selbst womöglich eine geeignete Krankheitsbezeichnung, und es bedarf zudem kontrollierter randomisierter Studien, die die kausale Bedeutung des vermuteten Sauerstoffmangels mittels nasaler Überdruckbeatmungstherapie evaluieren. Auch wenn an dieser Stelle die grundsätzlichen Probleme des unklaren klinischen Stellenwerts des Schlafapnoe-Syndroms (9) nicht weiter verfolgt werden können, soll aus didaktischen Gründen unser Beispiel weitergeführt werden.

Zur Frage der Verblindung der Studienevaluatoren ersehen wir aus der vorliegenden Publikation, dass der Vergleich zwischen Pulsoxymetrie und Polysomnographie in der Studie verblindet war.

– Schloss die Studie ein genügend breites Spektrum von Patienten ein, an welchen der Test in der klinischen Praxis angewendet wird?

Ein diagnostischer Test ist nur in dem Ausmass nützlich, als er zwischen gesuchten Zielcharakteristiken unterscheiden hilft, welche ansonsten nicht identifizierbar sind. Praktisch jeder diagnostische Test kann zwischen Schwerkranken und Gesunden unterscheiden. Der wahre diagnostische Wert einer Untersuchung wird deshalb nur in einer Studie erhellt, welche Patienten einschliesst, die die zu suchende Zielkrankheit in unterschiedlicher Ausprägung haben. Nur so kann evaluiert werden, wie sensitiv der Test leichtere Krankheitsstadien erfasst. Die Studie muss auch Patienten einschliessen, die

verschiedene Komorbiditäten aufweisen. Nur so kann evaluiert werden, wie spezifisch der Test die gesuchte Krankheit und nicht andere Störungen erfasst.

Die Studie schloss 114 Patienten ein, welche von Internisten und Spitälern mit der Verdachtsdiagnose eines Schlafapnoe-Syndroms der Pneumologischen Abteilung der Universitätsklinik Bonn zugewiesen wurden. Ob es sich um eine konsekutive und unselektionierte Stichprobe von Patienten handelt, ist unklar. Der durchschnittliche Bodymassindex (BMI) der Studienpatienten betrug $30,8 \pm 6,8 \text{ kg/m}^2$ (Mittelwert und Standardabweichung), und bei 75,9% waren anamnestic Apnoephasen beobachtet worden. Rund 73% der Patienten gaben Schlafneigungen an, und 95% litten unter Schnarchen. Alle Patienten hatten keine vorhergehende Schlafapnoeabklärungen. Angaben zum unterschiedlichen Ausmass und Dauer der typischen Symptome eines Schlafapnoe-Syndroms sowie dem Leidensdruck (verminderte Lebensqualität) fehlen hingegen. Es handelt sich also um eine etwas eng gefasste Patientengruppe mit ausgeprägten Symptomen, die einer Spezialabteilung zugewiesen wurde.

– Beeinflussten die Ergebnisse des zu evaluierenden Tests die Entscheidung, ob der Standardtest angewendet wurde?

Die Eigenschaften eines zu evaluierenden Tests werden verzerrt, falls dessen Ergebnisse die Entscheidung beeinflussen, ob Patienten mit dem Referenztest zusätzlich untersucht werden oder nicht. Diese Fehlermöglichkeit wird in der Literatur als «verification bias» (10) oder «work-up bias» (11) bezeichnet. Ein Beispiel soll dies verdeutlichen. In einer Studie zur Evaluation der Ventilations-Perfusions-Szintigraphie zur Diagnose der akuten Lungenembolie erhielten Patienten mit niedriger Wahrscheinlichkeit einer Lungenembolie im Szintigraphiebefund weniger häufig eine Angiographie (69%) als Patienten mit hochverdächtigem Szintigraphiebefund (92%) (12). Es ist verständlich, dass Kliniker wenig geneigt sind, Patienten weiteren invasiven Untersuchungen auszusetzen, wenn die Chance eines zusätzlichen Informationsgewinns gering ist. Die Autoren dieser Studie korrigierten jedoch diesen Bias,

indem sie Patienten mit normalen Szintigraphiebefunden über ein Jahr bezüglich des Auftretens von klinisch manifesten Lungenembolien nachkontrollierten.

Alle in die Studie eingeschlossenen Patienten, die eine nächtliche Oxymetrie erhielten, wurden einer Polysomnographie unterzogen. Ein «work-up bias» ist somit ausgeschlossen.

– Wurde die Testmethodik genügend detailliert beschrieben, um die Replikation des Tests zu ermöglichen?

Studien zu diagnostischen Tests sollten die Testmethoden exakt beschreiben, um deren Replizierbarkeit zu ermöglichen und um überprüfen zu können, ob der Test unter klinischen Alltagsbedingungen anwendbar ist. Wichtige Angaben zu Technik, Art der verwendeten Geräte oder Labormethoden, Patientenvorbereitung (Diät, Medikation unter der Testsituation etc.) sollten erwähnt sein.

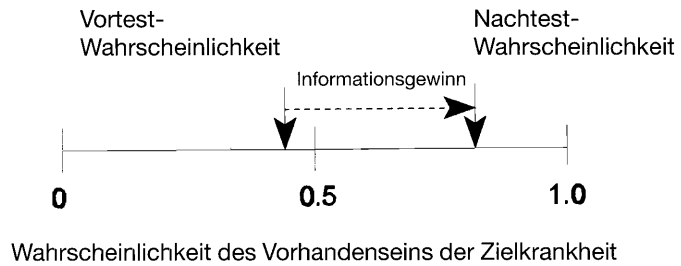
Die in die Studie eingeschlossenen Patienten erhielten alle eine ambulante Pulsoxymetrie. Zusätzlich wurden Schnarchgeräusche registriert sowie die Körperposition, unter welcher Apnoephasen auftraten. Eine Apnoe oder Hypnoe wurde als eine 4%-O₂-Entsättigungsabnahme definiert, und es wurden verschiedene Grenzwerte von ≥ 5 , ≥ 10 , ≥ 15 , ≥ 20 und ≥ 25 Entsättigungen pro Stunde evaluiert. Die Messparameter und Methoden der Polysomnographie werden detailliert aufgeführt, und die verwendeten Definitionen von Apnoe- und Schlafphasen sind angegeben.

2. Was sind die Studienergebnisse?

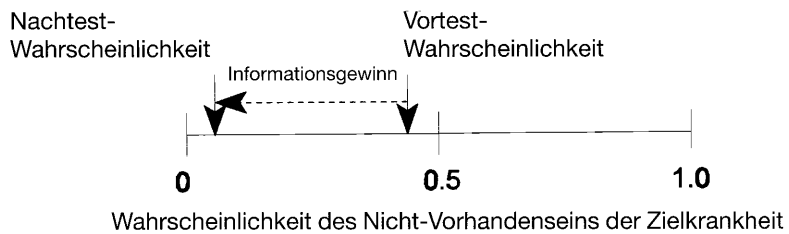
– Werden «Likelihood ratios» bzw. Sensitivität und Spezifität der Testergebnisse angegeben oder die Daten zu deren Berechnung aufgelistet?

Wenn wir einen diagnostischen Test anwenden, gehen wir von einem klinischen Verdachtsmoment aus, das heißt, wir gehen von einer bestimmten Wahrscheinlichkeit (oder Vortest-Wahrscheinlichkeit) aus, dass die gesuchte Zielkrankheit vorliegt (Abbildung 1). Wir verwenden einen Test mit Qualitäten, der uns mit hoher Wahrscheinlichkeit erlauben soll, eine Zielkrankheit möglichst sicher zu bestätigen oder auszuschließen, oder anders formuliert, der

Positives Testergebnis



Negatives Testergebnis



Veränderung der Vortest-Wahrscheinlichkeit durch den Informationsgewinn eines Tests. Wahrscheinlichkeiten werden mit Werten zwischen 0 und 1 ausgedrückt, der Wert 0 bedeutet kein Vorliegen und der Wert 1 sicheres (100%iges) Vorliegen der Zielkrankheit.

Abb. 1. Vortest- und Nachtest-Wahrscheinlichkeit bei einer Testuntersuchung

eine hohe oder niedrige Nachtest-Wahrscheinlichkeit für das Vorliegen einer Zielkrankheit erzeugt. Diese Qualität des Tests wird anhand der Sensitivität und der Spezifität bemessen. Ein gutes Konzept zur Beurteilung von Testqualitäten sind auch Wahrscheinlichkeitsraten, sogenannte «Likelihood ratios».

«Likelihood ratios» fassen die durch Sensitivität und Spezifität gegebene Testleistung in einer Zahl zusammen. Die «Likelihood ratio» für ein positives Testergebnis ist die Rate (oder das Verhältnis) der richtig positiven Testergebnisse, zu den falsch positiven Testergebnissen (Tabelle 2). Die Rate der richtig

Tab. 2. Vergleich der Ergebnisse eines diagnostischen Tests mit einem Standardtest

Ergebnis des Tests unter Evaluation	Referenztest (Standardtest)		
	Krankheit vorhanden	Krankheit nicht vorhanden	
Krankheit vorhanden	Richtig Positive a	Falsch Positive b	a+b
Krankheit nicht vorhanden	Falsch Negative c	Richtig Negative d	c+d
	a+c	b+d	a+b+c+d

Sensitivität = $a / (a+c)$

Spezifität = $d / (b+d)$

Likelihood ratio für ein positives Testergebnis = $[a/(a+c)]/[b/(b+d)] = \text{Sensitivität} / (1 - \text{Spezifität})$

Likelihood ratio für ein negatives Testergebnis = $[c/(a+c)]/[d/(b+d)] = (1 - \text{Sensitivität}) / \text{Spezifität}$

Positiver prädiktiver Wert = $a / (a+b)$

Negativer prädiktiver Wert = $d / (c+d)$

Prävalenz = $a+c / (a+b+c+d)$

Vortest-odds = Prävalenz / (1 - Prävalenz)

Nachtest-odds = Vortest-odds \times Likelihood ratio

Nachtest-Wahrscheinlichkeit = Nachtest-odds / (Nachtest-odds + 1)

positiven Testergebnisse gemessen am Total aller Patienten mit der gesuchten Zielkrankheit, entspricht der Sensitivität eines Tests ($a/a+c$). Die Rate der falsch positiven Testergebnisse ($b/b+d$) entspricht der Differenz $100\% - \text{Spezifität}$. Die «Likelihood ratio» für ein negatives Testergebnis ist als die Rate (oder das Verhältnis) der falsch negativen Testergebnisse zu den richtig negativen Testergebnissen definiert (Tabelle 2). Die Rate der falsch negativen Testergebnisse ($c/a+c$) entspricht der Differenz $100\% - \text{Sensitivität}$. Die Rate der richtig negativen Testergebnisse, gemessen am Total aller Patienten, die die gesuchte Krankheit nicht aufweisen, entspricht der Spezifität eines Tests ($d/b+d$).

Ein Vorteil der «Likelihood ratios» ist, dass sich aus der Vortest-Wahrscheinlichkeit mittels des Nomogramms von Fagan (13) unmittelbar die Nachtest-Wahrscheinlichkeit eines Testergebnisses berechnen lässt. Die Vortest-Wahrscheinlichkeit wird auch als die Prävalenz der Zielkrankheit bezeichnet.

Kehren wir zu unserem Beispiel zurück. Aufgrund Ihrer Beurteilung haben Sie die Wahrscheinlichkeit für das Vorliegen eines Schlafapnoe-Syndroms bei Ihrem Patienten auf rund 30% eingeschätzt. In Tabelle 3 sind die Sensitivität und Spezifität der ambulanten Pulsoxymetrie in Abhängigkeit verschiedener

Grenzwerte des O_2 -Entsättigungsindex aus der Originalpublikation aufgeführt. Bei einem Grenzwert von ≥ 10 Entsättigungen pro Stunde in der Puls-oxymetrie ergibt sich, wie die einfachen Berechnungen in Tabelle 3 zeigen, eine «Likelihood ratio» von 1.6 für einen positiven respektive von 0.12 für einen negativen Test. Eine «Likelihood ratio» von 1.6 besagt, dass eine positive Puls-oxymetrie bei einem Grenzwert von ≥ 10 Entsättigungen pro Stunde mit 1.6-facher grösserer Wahrscheinlichkeit bei einem Patienten vorkommt, bei dem in der Polysomnographie ein Schlafapnoe-Syndrom nachgewiesen wurde, als bei einem Patienten ohne Schlafapnoe-Syndrom. Entsprechend besagt eine «Likelihood ratio» von 0.12, dass ein solches Ergebnis viel wahrscheinlicher bei einem Patienten, bei dem sich in der Polysomnographie kein Schlafapnoe-Syndrom nachweisen lässt, zu erwarten ist. Je niedriger die Rate, um so höher ist die Wahrscheinlichkeit, dass es sich bei einem negativen Test handelt. Mit Hilfe des Nomogramms von Fagan (Abbildung 2) können die Nachtest-Wahrscheinlichkeiten einfach abgelesen werden. Wir suchen die Vortest-Wahrscheinlichkeit auf der linken Skala auf (z.B. Ihre Schätzung von 30%), ziehen eine Linie zur mittleren Skala der «Likelihood ratio» (in un-

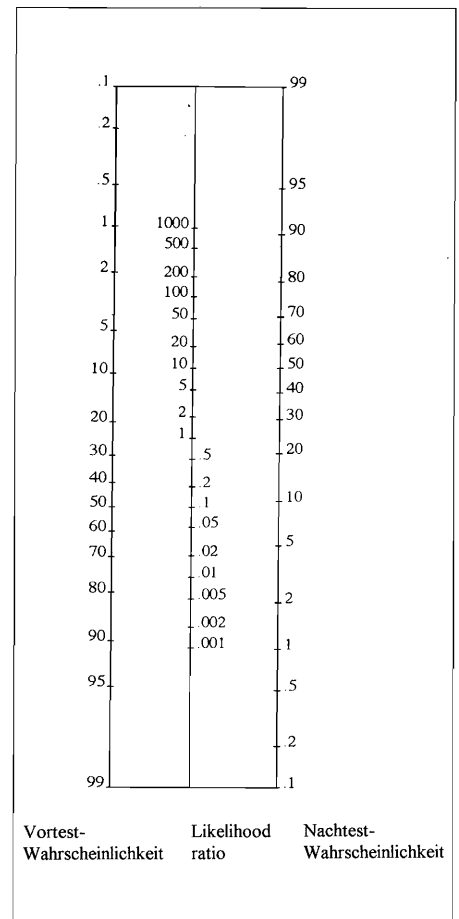


Abb. 2. «Likelihood ratio»-Nomogramm von Fagan zur Berechnung von Nachtest-Wahrscheinlichkeiten (Referenz 13)

Tab. 3. Sensitivität, Spezifität, positiver und negativer prädiktiver Wert sowie «Likelihood ratios» für unterschiedliche Grenzwerte des O_2 -Entsättigungsindex der ambulanten Pulsoxymetrie zur Diagnose eines Schlafapnoe-Syndroms (Referenz 5)

Ambulante Pulsoxymetrie Grenzwerte der O_2 -Entsättigung pro Stunde	Sensitivität %	Spezifität %	Positiver prädiktiver Wert %	Negativer prädiktiver Wert %	Likelihood ratio für positives Testergebnis	Likelihood ratio für negatives Testergebnis
≥ 5	96	15	73	63	1.1	0.26
≥ 10	95	41	79	78	1.6	0.12
≥ 15	83	62	84	60	2.2	0.27
≥ 20	68	74	86	49	3.1	0.43
≥ 25	60	85	91	48	4.0	0.47

Berechnungsbeispiel der Likelihood ratio für einen Grenzwert der O_2 -Entsättigung von ≥ 10 :

$$LR + = \text{Sensitivität} / (1 - \text{Spezifität}) = 95\% / (100\% - 41\%) = 1.6$$

$$LR - = (1 - \text{Sensitivität}) / \text{Spezifität} = (100\% - 95\%) / 41\% = 0.12$$

Berechnung der Nachtest-Wahrscheinlichkeit eines Schlafapnoe-Syndroms (Grenzwert der O_2 -Entsättigung ≥ 10 pro Stunde) bei einer Vortestwahrscheinlichkeit von 30% und positivem Test:

$$\text{Vortest-odds} = \text{Prävalenz} / (1 - \text{Prävalenz}) = 30\% / (100\% - 30\%) = 0.43$$

$$\text{Vortest-odds} \times \text{likelihood ratio} = \text{Nachtest-odds}; 0.43 \times 1.6 = 0.688$$

$$\text{Nachtest-Wahrscheinlichkeit} = [\text{Nachtest-odds} / (\text{Nachtest-odds} + 1)] \times 100 = [0.688 / (0.688 + 1)] \times 100 = 41\%$$

Berechnung der Nachtest-Wahrscheinlichkeit eines Schlafapnoe-Syndroms (Grenzwert der O_2 -Entsättigung ≥ 10 pro Stunde) bei einer Vortest-Wahrscheinlichkeit von 30% und negativem Test:

$$\text{Vortest odds} = \text{Prävalenz} / (1 - \text{Prävalenz}) = 30\% / (100\% - 30\%) = 0.43$$

$$\text{Vortest odds} \times \text{Likelihood ratio} = \text{Nachtest odds}; 0.43 \times 0.12 = 0.052$$

$$\text{Nachtest-Wahrscheinlichkeit} = [\text{Nachtest odds} / (\text{Nachtest odds} + 1)] \times 100 = [0.052 / (0.052 + 1)] \times 100 = 5\%$$

serem Beispiel 1.6 für einen positiven Test und 0.12 für einen negativen Test in der Pulsoxymetrie [Grenzwert von ≥ 10 Entsättigungen pro Stunde] und lesen auf der rechten Skala die Nachtest-Wahrscheinlichkeit ab. In unserem Beispiel beträgt die Nachtest-Wahrscheinlichkeit für einen positiven Test 41% und für einen negativen Test 5%. Würden wir die Einschätzung des Spezialisten übernehmen (Vortest-Wahrscheinlichkeit 50%), ersehen wir auf dem Nomogramm eine Nachtest-Wahrscheinlichkeit von 62% für einen positiven Test und eine solche von 11% für einen negativen Test in der Pulsoxymetrie.

Unser Beispiel macht deutlich, wie die Testausbeute von der Vortest-Wahrscheinlichkeit abhängt. Die Aussagekraft eines Tests hängt entscheidend von der Vortest-Wahrscheinlichkeit ab. Je höher bei gegebener Sensitivität und Spezifität die Vortest-Wahrscheinlichkeit, um so höher die prädiktive Wertigkeit des Tests. Die Nachtest-Wahrscheinlichkeit entspricht dem positiven respektive negativen prädiktiven Wert eines Tests. Der positive prädiktive Wert eines Tests entspricht der Rate der richtig positiven Testergebnisse an der Gesamtzahl aller positiven Testergebnisse (Tabelle 2: $[a/a+b]$). Dementsprechend entspricht der negative prädiktive Wert eines Tests der Rate der richtig negativen Testergebnisse an der Gesamtzahl aller negativen Testergebnisse (Tabelle 2: $[d/c+d]$). Allerdings ist die Spezifität der ambulanten Pulsoxymetrie sehr unbefriedigend, weswegen die positive prädiktive Wertigkeit der ambulanten Pulsoxymetrie gering ist und die Zunahme der «Likelihood ratios» bei grösseren Vortest-Wahrscheinlichkeiten ebenfalls gering ist.

Wie sich aus der Vortest-Wahrscheinlichkeit mittels dem Nomogramm von Fagan die Nachtest-Wahrscheinlichkeit eines Testergebnisses berechnen lässt, wird für den/die interessierte Leser/in nochmals im Appendix erläutert. Ein Nachteil von «Likelihood ratios» ist vielleicht, dass die Berechnung der Nachtest-Wahrscheinlichkeiten nur indirekt mit Hilfe des Nomogramms möglich ist. «Likelihood ratios» haben jedoch weitere Vorteile. So erlauben sie eine flexible Testinterpretation, wo sie für verschiedene gewählte Grenzwerte bekannt sind. Aufgrund der Vortest-Wahrscheinlichkeit kann beurteilt wer-

den, welcher Grenzwert gewählt werden kann, um auf eine klinisch relevante Nachtest-Wahrscheinlichkeit zu kommen. Tabelle 3 verdeutlicht, wie bei zunehmend höherem Grenzwert der O_2 -Entsättigung in der ambulanten Pulsoxymetrie die «Likelihood ratios» als auch der positive prädiktive Wert des Tests zunehmen. Umgekehrt zeigt sich, dass bei niederen Grenzwerten der O_2 -Entsättigung die «Likelihood ratios» für einen negativen Test abnehmen und die negative prädiktive Wertigkeit des Tests leichtgradig höher ist als bei höheren O_2 -Entsättigungsgrenzwerten.

Tests mit grossen «Likelihood ratios» erhöhen – bei gegebener Vortest-Wahrscheinlichkeit – die Wahrscheinlichkeit, dass die gesuchte Krankheit vorhanden ist. Umgekehrt machen Tests mit einer niedrigen «Likelihood ratio» das Vorliegen einer gesuchten Krankheit unwahrscheinlich. «Likelihood ratios» in der Grössenordnung von > 10 bei positivem Test oder < 0.1 bei negativem führen in der Regel zu Nachtest-Wahrscheinlichkeiten, die eine gesuchte Zielkrankheit sicher ein- respektive ausschliessen lassen. «Likelihood ratios» zwischen 5 und 10 respektive 0.1 und 0.2 ergeben intermediäre Veränderungen der Nachtest-Wahrscheinlichkeit. «Likelihood ratios» zwischen 2 und 5 respektive 0.5 und 0.2 bewirken geringe, jedoch manchmal wichtige Veränderungen der Nachtest-Wahrscheinlichkeit. «Likelihood ratios» zwischen 1 und 2 und 0.5 und 1 verändern die Nachtest-Wahrscheinlichkeit in einem klinisch kaum relevanten Ausmass.

Unser Beispiel zeigt, dass die «Likelihood ratios» der Pulsoxymetrie ungenügend sind, um in der Diagnose eines Schlafapnoe-Syndroms weiterzuhelfen. Hingegen ist bei einem Grenzwert von < 10 Entsättigungen pro Stunde die Sensitivität (95%) der Pulsoxymetrie recht gut. Bei einem Test mit einer hohen Sensitivität lässt sich bei einem negativen Testergebnis eine gesuchte Zielkrankheit mit grosser Wahrscheinlichkeit ausschliessen. Bei Vortestwahrscheinlichkeiten von 30% bis 50% und einem negativen Test < 10 Entsättigungen) resultieren Nachtest-Wahrscheinlichkeiten für das Vorliegen eines Schlafapnoe-Syndroms von unter 10%. Bei einem Grenzwert von < 10 Entsättigungen pro Stunde weist die Pulsoxymetrie eine nur ungenügen-

de Spezifität von 41% auf. Die Autoren haben deshalb überprüft, ob sich durch einen Score von klinischen Symptomen und den Ergebnissen der ambulanten Pulsoxymetrie die prädiktive Wertigkeit des Tests verbessern liesse. Bei einem Grenzwert von < 10 Entsättigungen pro Stunde, einem BMI von 29 kg/m^2 , dem Vorliegen von beobachteten Apnoephasen sowie unfreiwilliger Einschlafneigung fand sich eine Spezifität von 91% mit einem positiven prädiktiven Wert von 92% und einer Sensitivität von 41%. Die «Likelihood ratio» für einen positiven Test verbessert sich von 1.6 (Tabelle 3) auf immerhin 4.6 (Berechnung: $41\% / (100\% - 91\%)$). Bei einer Vortest-Wahrscheinlichkeit von z.B. 50% ergibt sich im Nomogramm eine Nachtest-Wahrscheinlichkeit von 70%, was einer erheblichen Verbesserung der Testausbeute entspricht. Bei einem Test mit einer hohen Spezifität lässt sich also bei einem positiven Testbefund die Zielkrankheit mit grosser Wahrscheinlichkeit einschliessen.

Vereinfachend können wir auch sagen, dass die Testausbeute von zwei Faktoren bestimmt wird, der «Likelihood ratio» (als zusammenfassendes Mass von Spezifität und Sensitivität) und der Vortest-Wahrscheinlichkeit. Je höher respektive niedriger die «Likelihood ratio», um so besser ist die Testausbeute. Der zweite entscheidende Faktor ist die Vortest-Wahrscheinlichkeit: Bei einer Vortest-Wahrscheinlichkeit zwischen 40% und 60% und gegebener Sensitivität und Spezifität ist die Testausbeute am höchsten, d.h. es werden die besten Differenzen zwischen Vortest- und Nachtest-Wahrscheinlichkeiten erzielt. Ein weiterer Vorteil von «Likelihood ratios» ist, dass die Ausbeute bei Anordnung von sequentiellen Tests analysiert werden kann. Hier geht es um Fragen wie z.B., was bringt eine Radionuklidventrikulographie an Zusatzinformation zu einem grenzwertigen Befund im Belastungs-EKG. In diesem Fall entspricht die «Nachtest-Wahrscheinlichkeit» des Belastungs-EKGs der «Vortest-Wahrscheinlichkeit» der Radionuklidventrikulographie. Die Bedeutung der «Likelihood ratios» zur Optimierung der klinischen Entscheidungsfindung ist in Zunahme begriffen. Es besteht bereits eine ausgezeichnete Sammlung zu Testevaluationen mit «Likelihood ratios» für häufige klinische Probleme (14).

3. Sind die Ergebnisse für die Behandlung meiner Patienten nützlich?

– Sind die Reproduzierbarkeit und die Interpretation der Testergebnisse in meinem klinischen Umfeld gegeben?

Eine wichtige Eigenschaft eines diagnostischen Tests, die es zu kennen gilt, ist dessen Reproduzierbarkeit, d.h. die Eigenschaft, z.B. bei stabilen Patienten oder Testkonditionen die gleichen Ergebnisse zu liefern. Eine schlechte Reproduzierbarkeit kann mit technischen oder methodischen Problemen eines Tests (z.B. Radioimmuno Assay) zusammenhängen. Eine unterschiedliche Testinterpretation von Untersuchern kann ebenso die Reproduzierbarkeit eines Tests beeinflussen. Deshalb sollten Angaben zur Reproduzierbarkeit vorhanden sein. Dies ist besonders dann wichtig, wenn die Testinterpretation spezielle Fähigkeiten der Untersucher erfordert (z.B. Interpretation von Computertomographien, Ultraschalluntersuchungen, Elektrokardiogramm etc.). Wenn die Reproduzierbarkeit eines Tests nur mittelmässig ist und der Test dennoch zwischen pathologischen und normalen Befunden unterscheiden hilft, besteht weniger Anlass zur Sorge. Ist die Reproduzierbarkeit eines Tests sehr hoch und variiert die Beurteilung durch die Untersucher nur geringgradig, dann ist der Test in seiner Anwendung entweder einfach, oder die Testinterpretation sind hoch qualifiziert. Falls letzteres zutrifft, kann der Test in einem anderen klinischen Umfeld eine geringere Reproduzierbarkeit haben.

Angaben zur Reproduzierbarkeit fehlen gänzlich. Im konkreten Fall muss gegebenenfalls mittels Nachfrage vor Ort überprüft werden, ob Qualitätsstandards oder Daten vorhanden sind, welche eine Beurteilung der Reproduzierbarkeit von diagnostischen Tests in unserem eigenen klinischen Umfeld zulassen.

– Können die Ergebnisse an meinem Patienten angewendet werden?

Bei dieser Frage geht es um die Testgenauigkeit. Testeigenschaften können sich ändern, wenn diagnostische Tests in unterschiedlichen Patientenpopulationen angewendet werden, die sich z.B. im Schweregrad der gesuchten Zielkrankheit unterscheiden. Ebenso können Störfaktoren wie z.B. Komorbidität,

welche das Testergebnis beeinflussen, variieren. Wenn die zu untersuchende Patientenpopulation kränker ist, d.h. eine ausgeprägtere Form der gesuchten Zielkrankheit hat, ergibt sich eine höhere Testsensitivität. Dies ist gleichbedeutend mit einer Zunahme der «Likelihood ratios», d.h. die «Likelihood ratios» liegen weiter von 1 entfernt. Falls die Patienten nur leichte Affektionen der Zielkrankheit aufweisen, bewegen sich die «Likelihood ratios» näher zu 1 (die Sensitivität ist geringer als aus einer kränkeren Studienpopulation vorgegeben). Falls Patienten ohne die gesuchte Zielkrankheit Faktoren aufweisen, die ein falsch positives Ergebnis begünstigen, dann bewegen sich die «Likelihood ratios» ebenfalls näher zu 1 und die Aussagekraft des Tests nimmt wiederum ab. Umgekehrt, falls Faktoren, welche das Testergebnis verfälschen können, in der Zielpopulation weniger häufig sind, nehmen die «Likelihood ratios» ab und die Aussagekraft des Tests steigt. Dass Testeigenschaften sich in Zielpopulationen, welche unterschiedliche Erkrankungsausprägungen aufweisen, ändern, ist gut dokumentiert.

Beispielsweise fanden sich bei Patienten mit ausgeprägteren Koronarstenosen bei positiven Belastungs-EKGs höhere «Likelihood ratios» (15). Allerdings ist die klinische Bedeutung dieser «Likelihood ratio»-Variation zweitrangig. Für die klinische Entscheidungsfindung viel wesentlicher bleibt die Berücksichtigung der Vortest-Wahrscheinlichkeit und deren Einfluss auf die Testaussagekraft. Sox drückt dies in seinem Standardwerk «Medical decision making» so aus: «If you once accept this principle, life will never be the same again» (16). *Genauere Angaben über die Art und Anzahl der Klinikzuweisungen fehlen in den vorliegenden Arbeiten. Der Vergleich mit Ihrem Patienten zeigt, dass Ihr Patient weniger Risikofaktoren und auch weniger ausgeprägte Symptome und einen niedrigeren BMI aufweist als der Durchschnitt der Studienpatienten. Sie schliessen hieraus, dass Ihr Patient nur bedingt mit der hier untersuchten Studienpopulation vergleichbar ist und dass die Testsensitivität im Vergleich zur Studienpopulation bei Ihrem Patienten geringer sein kann.*

– Werden die Testergebnisse das Patientenmanagement beeinflussen?

Unsere Ausführungen sollen zeigen, in-

wiefern explizite Überlegungen zur Wahrscheinlichkeit des Vorhandenseins einer gesuchten Zielkrankheit die klinische Entscheidungsfindung bei diagnostischen Problemen leiten können. Es ist wichtig, dass für jede Zielkrankheit Schwellenwerte definiert werden, deren Unter- beziehungsweise Überschreitung keine weitere Abklärungen beziehungsweise eine Weiterabklärung oder direkte Behandlung nach sich ziehen.

Auflösung des klinischen Szenarios

Ausgerüstet mit dieser Information machen Sie sich an die Lösung Ihres klinischen Problems. Aus einer nochmaligen Befragung des Patienten erfahren Sie, dass er nach Spätschichten wegen Einschlafproblemen spät nachts Alkohol trinkt und dann tagsüber Einschlafneigungen hat. Während Tagschichten ist der Patient kaum schläfrig, und die Partnerin beobachtet, dass Ihr Patient nachts weniger schnarcht. Apnoephasen habe sie nicht beobachtet. Aufgrund dieser Angaben und der Tatsache, dass der Patient nur mässig übergewichtig ist, schätzen Sie die Vortest-Wahrscheinlichkeit eines Schlafapnoe-Syndroms bei Ihrem Patienten auf maximal 30%. Mit dem Nomogramm berechnen Sie bei einem Grenzwert von ≥ 10 Entsättigungen pro Stunde eine Nachttest-Wahrscheinlichkeit von 41%. Dieser Informationsgewinn ist klinisch unbedeutend. Wenn es Ihr Ziel ist, ein Schlafapnoe-Syndrom bei Ihrem Patienten auszuschliessen, sieht die Rechnung so aus: Bei einem Grenzwert von ≥ 15 Entsättigungen pro Stunde ergibt sich eine «Likelihood ratio» für einen negativen Test von 0.27 (Berechnung: $[100\% - 83\%] / 62\% = 0.27$). Bei einer Vortest-Wahrscheinlichkeit von 30% ersehen wir aus dem Nomogramm eine Nachttest-Wahrscheinlichkeit von 10%. Damit kann der Test uns auch nicht helfen, ein Schlafapnoe-Syndrom sicher auszuschliessen, auch wenn das vorliegende Ergebnis darauf hinweist, dass die Wahrscheinlichkeit für das Vorliegen eines Schlafapnoe-Syndroms eher als gering zu betrachten ist. Sie raten Ihrem Patienten vorläufig von einer Schlafabklärung ab. Sie instruieren den Patienten bezüglich Schlafhygiene und vereinbaren eine Reduktion des Alkoholkonsums. Ihr Patient ist einverstanden, und eine Verlaufsbeobachtung sowie eine anschliessende nochmalige

Befragung, die keine weitere Indizien auf ein Schlafapnoe-Syndrom ergeben, bestätigen Ihre Einschätzung.

Appendix

Nachfolgend zeigen wir die genauen Berechnungen der Vortest- und Nachtest-Wahrscheinlichkeiten respektive der «Vortest- und Nachtest-odds». Die mathematisch korrekte Definition der Vortest-Wahrscheinlichkeit ist der englische Begriff «pretest odds», der im Deutschen am ehesten mit dem Begriff «Vortest-Ereignisance» übersetzt werden kann und mit folgender Formel ausgedrückt wird: Prävalenz/(1-Prävalenz) (Tabelle 2). Diese Rate besagt nichts anderes als die Wahrscheinlichkeit eines Ereigniseintritts gegenüber der Wahrscheinlichkeit des Nichtauftretens eines Ereignisses. In unserem Beispiel gehen wir von einem Grenzwert der ambulanten Pulsoxymetrie von 10 Entsättigungen pro Stunde (Tabelle 3) und einer Vortest-Wahrscheinlichkeit von 30% für das Vorliegen eines Schlafapnoe-Syndroms aus. Die «Vortest-odds» beträgt $30\% / (100\% - 30\%) = 0.43$. Das Konzept der «odds» ist im Wettsport bestens bekannt. Wir sagen z.B., die Chance, dass ein bestimmtes Pferd gewinnt, ist 1:5. Für uns Kliniker ist dieses Konzept vorerst befremdend, sind wir doch gewohnt, Vortest-Wahrscheinlichkeiten in Begriffen wie «wahrscheinlich» versus «nicht wahrscheinlich» oder eventuell in Prozentwerten auszudrücken. Die «Nachtest-Odds» berechnet sich nach folgender einfacher Formel: «Vor-

test-Odds» \times «Likelihood ratio» = «Nachtest-Odds». In unserem Beispiel ergibt sich also für ein positives Testergebnis eine «Nachtest-Odds» von $0.43 \times 1.6 = 0.688$. Die «Nachtest-Odds» muss dann in eine Nachtest-Wahrscheinlichkeit in Prozentwerten umgerechnet werden. Die Formel zur Umrechnung ist die folgende: «Nachtest-Odds» / («Nachtest-Odds» + 1). In unserem Beispiel ergibt sich eine Nachtest-Wahrscheinlichkeit für einen positiven Test von $0.688 / (0.688 + 1) = 40.8\%$. Die «Nachtest-Odds» für ein negatives Testergebnis beträgt $0.43 \times 0.12 = 0.05$, und die Nachtest-Wahrscheinlichkeit bei einem negativen Test für das Vorliegen eines Schlafapnoe-Syndroms beträgt $0.05 / (0.05 + 1) = 5\%$. □

Wir bedanken uns bei Dr. Daniel Pewsner, Bern, für die kritische Durchsicht des Manuskriptes und für Verbesserungsvorschläge.

Bibliographie

1. Van Surell C., Lemaigre D., Leroy M., Foucher A., Hagenmuller M.P., Raffestin B.: Evaluation of an ambulatory device, CID 102, in the diagnosis of obstructive sleep apnoea syndrome. *Eur Respir J* 1995; 8: 795-800.
2. Zucconi M., Ferini-Strambi L., Castronovo V., Oldani A., Smirne S.: An unattended device for sleep-related breathing disorders: validation study in suspected obstructive sleep apnoea syndrome. *Eur Respir J* 1996; 9: 1251-6.
3. Sériès F., Marc I., Cormier Y., La Forge J.: Utility of nocturnal home oxymetry for case finding in patients with suspected sleep apnea syndrome. *Ann Intern Med* 1993; 119: 449-53.
4. Gyulay S., Olson L.G., King H.M.T., Allen M., Saunders N.A.: A comparison of clinical assessment and home oximetry in the diagnosis of obstructive sleep apnea. *Am Rev Rspir Dis* 1993; 147: 50-53.
5. Schäfer H., Ewig S., Hasper E., Lüderitz B.: Predictive diagnostic value of clinical assessment and nonlaboratory monitoring system recordings in patients with symptoms suggestive of obstructive sleep apnea syndrome. *Respiration* 1997; 64: 194-199.
6. Jaeschke R., Guyatt G.H., Sackett D.L. for the Evidence-Based Medicine Working Group. User's Guide to the Medical Literature. III How to Use an Article About a Diagnostic Test. A. Are the Results of the Study Valid? *JAMA* 1994; 271: 389-91.
7. Jaeschke R., Guyatt G.H., Sackett D.L. for the Evidence-Based Medicine Working Group. User's Guide to the Medical Literature. III How to Use an Article About a Diagnostic Test. B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA* 1994; 271: 703-7.
8. Harvey R.L., Roth E.J., Arnold P.R., Durham J.R., Green D.: Deep vein thrombosis in stroke. The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke* 1996; 27: 1516-20.
9. Wright J., Johns R., Watt I., Melville A., Sheldon T.: Health effects of obstructive sleep apnoea and the effectiveness of continuous positive airways pressure: a systematic review of the evidence. *BMJ* 1997; 314: 851-60.
10. Begg C.B., Greenes R.A.: Assment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; 39: 207-15.
11. Choi B.C.K.: Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992; 45: 581-6.
12. The PIOPED Investigators: Value of ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA* 1990; 263: 2753-9.
13. Fagan T.J.: Nomogram for Bayes's theorem (C). *N Engl J Med* 1975; 293: 257.
14. Panzer R.J., Black E.R., Griner P.F.: Diagnostic Strategies for Common Medical Problems. Philadelphia, Pa: American College of Physicians; 1991.
15. Hlatky M.A., Pryor D.B., Harrell F.E.: Factors affecting sensitivity and specificity of exercise cardiography. *Am J Med* 1984; 77: 64-71.
16. Sox H.C., Blatt M.A., Higgins M.C., Marton K.I.: Medical Decision Making. Butterworth-Heinemann, Boston 1988.