

EbM-Splitter 9

Die Vierfeldertafel (in Diagnosestudien): Sensitivität und Spezifität

Die Ergebnisse unterschiedlicher Studientypen (z.B. Interventionsstudie, Diagnosestudie) können häufig in einer sogenannten Vierfeldertafel wiedergegeben werden. Eine Vierfeldertafel ist eine Zusammenfassung zweier Merkmale mit jeweils zwei Ausprägungen; in der Tafel werden die absoluten Häufigkeiten des Auftretens der vier möglichen Kombinationen abgetragen. Obwohl diese Tafeln immer die gleiche Form haben, werden sie abhängig vom Studientyp unterschiedlich erzeugt; dementsprechend unterscheiden sich auch die Auswertungsmethoden. Im Folgenden wird auf die Bedeutung der Vierfeldertafel in Diagnosestudien eingegangen.

Diagnosestudien

Zweck einer Diagnosestudie (Synonym: Diagnostikstudie) ist die Beurteilung der Güte einer (neuen) Maßnahme zur Diagnose einer bestimmten Krankheit. Dazu wird die Vorhersage der neuen diagnostischen Maßnahme mit einem sogenannten Goldstandard verglichen. Ein Goldstandard ist gewöhnlich durch die gegenwärtig zuverlässigste Methode zur Diagnose der jeweiligen Krankheit gegeben. Sehr häufig wird das Ergebnis einer diagnostischen Maßnahme in einer einfachen binären Entscheidung zusammengefasst, nämlich, ob die Krankheit vorliegt oder nicht. In diesem Fall kann das Ergebnis eines diagnostischen Tests in einer Vierfeldertafel dargestellt werden (siehe Tab. 1).

		Goldstandard Krankheit		
		liegt vor	liegt nicht vor	
Neues Diagnose- verfahren	Krankheit liegt vor (positives Testergebnis)	<i>richtig positiv</i> a	<i>falsch positiv</i> b	alle Test-Positiven $a+b$
	Liegt nicht vor (negatives Testergebnis)	<i>falsch negativ</i> c	<i>richtig negativ</i> d	$c+d$ alle Test-Negativen
		$a+c$ alle Erkrankten	$b+d$ alle Gesunden	N alle Untersuchten

Tab. 1: Allgemeine Darstellung einer Vierfeldertafel in einer Diagnosestudie

Bei Diagnosestudien ist es zur Konvention geworden, die Spalten der Tafel anhand des Goldstandards und die Zeilen anhand des Ergebnisses der neuen diagnostischen Maßnahme einzuteilen. Eine weitere Konvention besteht darin, die einzelnen Zellen zeilenweise von a bis d zu bezeichnen. Der Wert a beispielsweise gibt die Anzahl von Personen mit der Krankheit wieder, die mit der neuen diagnostischen Maßnahme als erkrankt erkannt werden. Häufig wird die Gesamtanzahl von Beobachtungen, d.h. $N=a+b+c+d$, in die rechte untere Ecke geschrieben. Die Anzahlen a und d entsprechen richtigen Vorhersagen („richtig positiv“ bzw. „richtig negativ“) und sollten bei einer neuen diagnostischen Maßnahme möglichst groß sein.

“[...]“ *Aids-Test positiv* “. Das ist die **negativste** Auskunft, die ein Mensch heute erhalten kann, die Todesbotschaft – fahrlässig und jargongläubig ins Gewand

des **Positiven** gehüllt. Ja doch, medizinisch absolut korrekt! Nur ohne irgendeine Ahnung von Sprache und Psychologie.“

Der Sprachstilist und Sprachkritiker Wolf Schneider zu „positiven“ Ergebnissen nach Anwendung diagnostischer Tests [4]

Sensitivität und Spezifität

Zur Beurteilung der Qualität einer diagnostischen Maßnahme betrachtet man gewöhnlich getrennt für erkrankte und gesunde Personen – die Wahrscheinlichkeit für eine richtige Vorhersage mittels der neuen diagnostischen Maßnahme. Die Wahrscheinlichkeit, eine erkrankte Person als krank zu erkennen, wird als *Sensitivität* bezeichnet. Eine Schätzung der Sensitivität ergibt sich zu $a/(a+c)$. Die Wahrscheinlichkeit, eine gesunde Person als gesund zu erkennen, wird als *Spezifität* bezeichnet. Eine Schätzung der Spezifität ergibt sich zu $d/(b+d)$. Zur Berechnung von Konfidenzintervallen für Sensitivität und Spezifität kann auf eine Standardmethode zurückgegriffen werden [1].

Eine optimale diagnostische Maßnahme hat Werte von jeweils 100% für Sensitivität und Spezifität; in diesem Fall sind alle Vorhersagen richtig. Zum Vergleich zweier diagnostischer Maßnahmen können die Werte für Sensitivität und Spezifität herangezogen werden. Ein Verfahren ist eindeutig vorzuziehen, falls es sich sowohl durch eine höhere Sensitivität als auch durch eine höhere Spezifität auszeichnet. Ist dies nicht der Fall, so ist eine differenziertere Abwägung notwendig, welche sich an den Konsequenzen einer Fehldiagnostik bei erkrankten bzw. gesunden Patienten orientieren muss.

Beispiel: Diagnostik der Approximalkaries

In einer an der Universität Erfurt [2] durchgeführten Studie zur Diagnostik der Approximalkaries wurde die Güte einer klinischen Untersuchung mit Spiegel und zahnärztlicher Sonde sowie mit Faseroptiktransillumination (FOTI) im Vergleich zu einer Bissflügelröntgenaufnahme (Goldstandard) bestimmt. Bei 186 Probanden wurde für insgesamt 5722 Approximalflächen jeweils eine Einteilung in kariös bzw. gesund (nicht-kariös) vorgenommen. Die Ergebnisse der beiden Verfahren im Vergleich zur röntgenologischen Untersuchung sind in Tabelle 2 und 3 zusammengefasst.

		Röntgenologische Untersuchung		S
		Kariös	Gesund	
Klinische Untersuchung	Kariös	172	5	177
	Gesund	385	5160	5545
S		557	5165	5722

Tab. 2: Vergleich der klinischen und röntgenologischen Kariesdiagnostik [2]

Mittels der klinischen Untersuchung wird ein gesunder Zahn mit großer Sicherheit auch als gesund erkannt: von insgesamt 5165 gesunden Approximalflächen ist dies bei immerhin 5160 der Fall, d.h. nur 5 Approximalflächen werden fälschlicherweise als kariös eingestuft (Tabelle 2). Hingegen ist die klinische Untersuchung nicht sonderlich gut dafür geeignet, eine kariöse Approximalfläche als solche zu erkennen: von 557 kariösen Approximalflächen werden nur 172 Approximalflächen als kariös eingestuft, d.h. der Großteil der kariösen

Approximalflächen wird nicht erkannt. Für die klinische Untersuchung ergibt sich somit eine Sensitivität von $172/557 = 30,9\%$ (95%-Konfidenzintervall: 27,0% bis 34,7%) und eine Spezifität von $5160/5165 = 99,9\%$ (95%-Konfidenzintervall: 99,8% bis 100%).

		Röntgenologische Untersuchung		S
		Kariös	Gesund	
Faseroptische Untersuchung	Kariös	377	26	403
	Gesund	180	5139	5319
	S	557	5165	5722

Tab. 3: Vergleich der faseroptischen und röntgenologischen Kariesdiagnostik [2]

Die Faseroptiktransillumination eignet sich im Vergleich zur klinischen Untersuchung erheblich besser, eine kariöse Approximalfläche als solche zu erkennen: von 557 kariösen Approximalflächen werden 377 Approximalflächen als kariös eingestuft (Tabelle 3). Die Sensitivität der faseroptischen Untersuchung beträgt somit $377/557 = 67,7\%$ (95%-Konfidenzintervall: 63,8% bis 71,6%). Die Wahrscheinlichkeit, eine gesunde Approximalfläche als gesund zu erkennen, ist mit Hilfe der faseroptischen Untersuchung ebenfalls sehr hoch; für die Spezifität ergibt sich ein Wert von $5139/5165 = 99,5\%$ (95%-Konfidenzintervall: 99,3% bis 99,7%). Aufgrund der erheblich höheren Sensitivität bei vergleichbarer Spezifität ist die FOTI der klinischen Untersuchung klar vorzuziehen. Dennoch kann die FOTI die röntgenologische Untersuchung nicht ersetzen. FOTI wird daher vor allem in Fällen empfohlen, in welchen die Anfertigung von Röntgenbildern nicht möglich ist; ferner kann sie als zusätzliches Hilfsmittel zur Sicherung der Diagnose herangezogen werden [3].

Schlussbemerkungen

Im Beispiel wurden die Ergebnisse der Bissflügelaufnahme als Goldstandard zur Erkennung einer Approximalkaries verwendet. Diese Vorgehensweise ist insoweit problematisch, als Unterschiede zwischen in situ erstellten Röntgenbefunden und den Befunden aus einer mikroradiographischen oder histologischen Untersuchung extrahierter Zähne zu erwarten sind. So wird die röntgenologisch diagnostizierte Ausdehnung einer Karies grundsätzlich unterschätzt [3]. Gleichwohl stellt die Bissflügelaufnahme unter den Verfahren, die keine Extraktion der Zähne benötigen (Mikroradiographie von Zahndünnschliffen; histologische Untersuchung der Zahnhartsubstanzen), die verlässlichste Methode dar.

Sensitivität und Spezifität sind zwei Kenngrößen, welche die Güte einer diagnostischen Maßnahme beschreiben. Anhand dieser Werte kann beurteilt werden, wie gut eine gesunde bzw. erkrankte Person mittels des Diagnoseverfahrens als solche erkannt wird. Über die Auswirkungen der Anwendung der diagnostischen Maßnahme in der Praxis, d.h. bei Vorliegen eines positiven bzw. negativen Testergebnisses, geben Sensitivität und Spezifität jedoch nur indirekt Auskunft. Hierzu sollten die prädiktiven Werte (positiver und negativer Vorhersagewert) verwendet werden, welche einem der folgenden EbM-Splitter behandelt werden.

*Guido Schwarzer, Freiburg
Jens C. Türp, Freiburg / Basel
Gerd Antes, Freiburg*

Literatur

1. *Altman, D.G.:* Practical Statistics for Medical Research. Chapman & Hall, London 1991, 210.
2. *Heinrich, R., Künzel, W., Tawfiq, H.:* Approximale Kariesdiagnostik – Vergleich klinischer, faseroptischer und röntgenographischer Diagnoseverfahren (Approximalkariesdiagnostik). Dtsch Zahn Mund Kieferheilkd Zentralbl 79, 535 (1991).
3. *Hellwig, E., Klimek, J., Attin, T.:* Einführung in die Zahnerhaltung. 2. Aufl. Urban & Fischer, München 1999, 62-65.
4. *Schneider, W.:* Deutsch für Kenner. Die neue Stilkunde. 5. Aufl. Piper, München 2000, 146.

Deutsche Zahnärztliche Zeitschrift 57: 333-334 (2002)