Einführung in die klinische Epidemiologie

Wolfgang Klee und Carola Sauter-Louis

Ausgabe 2020

Inhalt

1. Vorwort 3
2. Diagnostik und Diagnose 4
2.1 Krankheitsbegriff 4
2.2 Grundlagen der Klassifikation 6
2.3 Testtheorie 7
2.4 Nosographie 24
2.5 „Normalwerte“ und Referenzintervalle 28
2.6 Trennwerte 29
2.7 Receiver Operating Characteristics (ROC)-Kurven 34
2.8 Goldstandard 39
2.9 Das Bayes-Theorem 42
2.10 Übereinstimmung von Testergebnissen (Kappa-Statistik) 43
2.11 Diagnostik-Methoden 47
2.12 Diagnostik in Populationen 50
3. Therapie 64
3.1. Beurteilung der Wirksamkeit von Interventionen 64
3.2 Hypothesen-basierte Forschung 65
3.3 Klinische Entscheidungsanalyse 65
4. Prophylaxe 76
4.1. Kausalität in der Medizin 76
4.2 Maße für Häufigkeit von Krankheiten 79
4.3 Wahrscheinlichkeit 81
4.4 Studien zur Quantifizierung des Zusammenhangs zwischen Faktoren und Krankheiten 83
5. Subjektive Aspekte 89
5.1 Wahrnehmung 89
5.2 Informationsverarbeitung 89
5.3 (Klinische) Erfahrung 90
5.4 Fehldiagnosen 90
6. Antworten auf die Fragen und Übungsaufgaben im Text 92
7. Literaturhinweise 99
8. Stichwortverzeichnis 101

1. Vorwort

Die traditionelle (tier)medizinische Ausbildung beinhaltet einen naturwissenschaftlichen, einen anatomisch-physiologischen, einen „paraklinischen“ und einen klinischen Teil. Letzterer umfasst im Allgemeinen die klinische Propädeutik, die Nosographie und den Unterricht am Patienten.
In der *Propädeutik* werden die Lebensäußerungen des gesunden und kranken Organismus sowie Untersuchungstechniken gelehrt. Unter dem Aspekt der Diagnostik bedeutet das die Vermittlung von Kenntnissen und Fertigkeiten zur Erhebung von Befunden, welche in den Prozess der Diagnostik eingehen.
Die *Nosographie*, also die systematische Beschreibung von Krankheiten, wird meist in Form von Vorlesungen (Krankheit A: Symptome a, b, c, Therapie, Prophylaxe; Krankheit B: Symptome b, d, e ...) angeboten.
Beim *Unterricht am Patienten* (klinische Demonstrationen, „Quote“) werden Differentialdiagnostik (Patient zeigt Symptome x, y, z, ... Es kommen in Frage: Krankheit A, Krankheit B ...) und therapeutische Technik gelehrt.
Quantitative Aspekte des rationalen Umgangs mit klinisch relevanter Information und mit Unsicherheit werden jedoch meist nicht vermittelt, obwohl sie unverzichtbarer Teil der medizinischen Ausbildung sind. Diese Aspekte stellen den Hauptteil dieser Einführung in die klinische Epidemiologie dar, auch wenn unter Epidemiologie im herkömmlichen Sinn das Auffinden und die Bewertung von Faktoren, welche Auftreten, Verbreitung und Verlauf von Krankheiten beeinflussen, verstanden werden.

In dem Bestreben, einen Bezug zur klinischen Ausbildung herzustellen, haben wir das Skriptum in die großen Bereiche Diagnostik, Therapie und Prophylaxe gegliedert.

Ziel dieser Einführung ist es, bei Studierenden der Tiermedizin Interesse an quantitativen Aspekten der Medizin zu wecken. Auch wenn der berühmte Astrophysiker HAWKING warnte, dass man mit jeder Formel sein Auditorium halbiert, und das Interesse der Studierenden an diesem Gebiet erfahrungsgemäß ohnehin begrenzt ist (um es ganz vorsichtig auszudrücken), kommt dieses Skript nicht ganz ohne (einfache) Mathematik aus. Nicht behandelt werden anspruchsvollere mathematische Methoden. Hierzu wird auf die zahlreichen ausführlichen Bücher über Epidemiologie und Biomathematik hingewiesen, von denen einige in den Literaturhinweisen aufgeführt sind.

Ein gewisses Problem der Epidemiologie kann darin bestehen, dass sich anhand ausgesuchter (oder erfundener) Daten epidemiologische Prinzipien schön demonstrieren lassen, dass aber die Anwendung in der Realität schwierig bis unmöglich ist (oder aber die Verhältnisse offensichtlich sind).

In den Text sind einige Fragen eingebaut. Es wird empfohlen, sie nicht zu überspringen und sie zunächst ohne Hilfe der in Kapitel 7 angebotenen Antworten zu bearbeiten. Die Antworten enthalten zum Teil zusätzliche Informationen.

Lernziele: Sicherer Umgang mit den Begriffen Sensitivität, Spezifität, prädiktiver Wert; Bewertung von Testkombinationen; Durchführung von klinischen Entscheidungsanalysen; Auffindung und Bewertung relevanter Informationen aus der Fachliteratur; Auswahl adäquater Studien zur Bearbeitung epidemiologischer Fragestellungen.

2. Diagnostik und Diagnose

„Vor die Therapie haben die Götter die Diagnose gestellt." Dieser häufig zitierte Satz, der dem Schweizer Kliniker NAEGELI zugeschrieben wird, verdeutlicht die Bedeutung, welche in der Medizin der Diagnose beigemessen wird. Aber kaum einer spricht über die theoretischen Grundlagen und Probleme der Diagnostik.

Eine Diagnose ist die Zuordnung eines konkreten Erkrankungsfalles zu einer nosologischen Einheit (Krankheit) eines nosologischen Systems. Unter Diagnostik wird hier die zur Diagnose führende, im Wesentlichen geistige Tätigkeit verstanden, auch wenn in der Realität die klinische Untersuchung natürlich einen essentiellen Teil der Diagnostik ausmacht.

Der Krankheitsbegriff wird im Abschnitt 2.1 näher beleuchtet. Einige Aspekte der Klassifikation werden im Abschnitt 2.2 abgehandelt.

2.1 Krankheitsbegriff

Viele Begriffe, die wir im täglichen Leben mehr oder weniger unreflektiert verwenden, erweisen sich bei näherer Betrachtung als recht problematisch. So auch der Begriff der Krankheit, nicht im Sinne von „Nicht-Gesundheit", sondern im Sinne von Krankheitseinheit. Was bedeutet es, wenn wir sagen: es gibt verschiedene Krankheiten? Diese Aussage impliziert eine Art von Existenz von Krankheiten.

Man unterscheidet im Prinzip zwei Arten von Krankheitsbegriffen: den ontologischen oder substantiellen und den nominalistischen. Der *ontologische Krankheitsbegriff* wurde vor allem von dem englischen Arzt Thomas SYDENHAM („der englische Hippokrates", 1624 - 1689) propagiert. Für ihn waren Krankheiten Naturkonstanten, die bei einem Sokrates den gleichen Verlauf zeigen wie bei einem Simpel, wie er sich ausdrückte. Er verglich sie mit den Spezies im Pflanzen- und Tierreich und forderte die Erstellung einer analogen Systematik.
Dieser Krankheitsbegriff dürfte abgeleitet sein von der Vorstellung, dass sich Dämonen des Körpers oder Geistes eines Menschen oder eines Tieres bemächtigen und so eine Krankheit herbeiführen, ihn aber auch wieder verlassen können. Infektiöse und parasitäre Krankheiten und auch Vergiftungen haben Wesenszüge, welche dieser Vorstellung nahekommen. Für Stoffwechsel- und Mangelkrankheiten sowie für Erbkrankheiten ist dieser Begriff weniger passend. Für die Vertreter des *nominalistischen Krankheitsbegriffes* sind Krankheiten nur Namen (lat.: nomina), die von Menschen als Hilfsgrößen zur Orientierung geschaffen werden. Der Satz „Es gibt keine Krankheiten, sondern nur kranke Individuen" bringt diese Vorstellung zum Ausdruck. Die praktische Relevanz der Unterscheidung ist jedoch begrenzt.
Mitunter werden Krankheitseinheiten auf der Basis neuerer Forschungsergebnisse aufgeteilt". Beispielsweise werden „Pansentrinken" und „D-Laktat-Azidose", die es natürlich schon vor ihrer Entdeckung gegeben hat, seither aus Komplex „Kälberdurchfall" ausgegliedert. Mit der Entdeckung der Rota- und Coronaviren sowie der Wiederentdeckung der Kryptosporidien zeigte sich, dass die frühere Bezeichnung "Colibacillose" für den Kälberdurchfall nicht korrekt war.

Zwei modernere Autoren definieren Krankheiten auf folgende Weisen:
„A disease is the sum of the abnormal phenomena displayed by a group of living organisms in association with a specified common characteristic or a set of characteristics by which they

differ from the norm of their species in such a way as to place them at a biological disadvantage.“ (SCADDING, 1967). Eine Krankheit ist eine Gruppe von Symptomen mit einheitlicher Entstehung (Pathogenese) und einheitlicher tieferer Ursache (Ätiologie).“ (GROSS, 1969)

Man kann seine eigene, vielleicht unbewusste Einstellung zu diesem Thema leicht dadurch feststellen, dass man sich fragt, ob Krankheiten entdeckt werden können. Das ist nämlich nur bei Verwendung des ontologischen Krankheitsbegriffes möglich.

Zweifellos können Krankheiten neu auftreten. Als Beispiel sollen die so genannte Hyänenkrankheit, die Mitte der 70er Jahre des vorigen Jahrhunderts zunächst in Frankreich und danach auch in anderen Ländern auftrat, und die bovine neonatale Panzytopenie (BNP), der in verschiedenen Ländern Europas Tausende von Kälbern zum Opfer fielen, genannt werden (s. Abb. 2.1-1 und (Abb. 2.1-2).

Abb. 2.1-1: Rotbuntes Jungrind mit der so genannten Hyänenkrankheit

Abb. 2.1-2: Petechien bei einem Kalb mit BNP

2.2 Grundlagen der Klassifikation

In der am Anfang gegebenen ausführlichen Definition des Begriffes "Diagnose" tauchte der Begriff "nosologisches System" auf. Man könnte das auch eine Systematik oder Klassifikation oder Taxonomie von Krankheiten nennen. Versuche, solche Klassifikationen zu erstellen, hat es schon seit langer Zeit gegeben. Thomas SYDENHAM hat das sehr ausdrücklich empfohlen und die Analogie zum System der Botanik betont. Von LINNÉ, der große biologische Systematiker, hat sich auch an einer Krankheiten-Systematik versucht.

Eine Klassifikation muss so beschaffen sein, dass alle möglichen Fälle einer und nur einer Kategorie zugeordnet werden können. Die logischen oder formalen Forderungen an eine Klassifikation bestehen also in *Vollständigkeit und Eindeutigkeit*. Die Tatsachen, dass ein Individuum zu einem bestimmten Zeitpunkt mehr als eine Krankheit haben kann, und dass es zu einer Symptomenkombination eine mehr oder weniger lange Liste von Differentialdiagnosen also in Frage kommende Krankheitseinheiten, geben kann, zeigen, dass diese Forderungen in der Medizin nicht immer (leicht) zu erfüllen sind.

Neben den beiden genannten logischen Forderungen an eine Systematik oder Klassifikation gibt es aber noch zwei mehr praktische: die *Zweckmäßigkeit und die „Praxisnähe"*. Unter dem Aspekt der Zweckmäßigkeit sind vor allem zwei Fragen von Bedeutung:

1. Was will ich wissen? (Die Antwort darauf legt Anzahl und Namen der Kategorien fest.)
2. Welches Irrtumsrisiko toleriere ich bei der Zuordnung? (Die Antwort darauf legt die Grenzwerte bei den Zuordnungskriterien fest.).

Die praktische Medizin als handlungsorientierte Disziplin braucht Kategorien, zu denen eine Zuordnung auf der Basis klinischer Daten möglich ist, und aus denen Handlungsanweisungen oder –legitimation ableitbar sind.
Jede Kategorisierung setzt die Ziehung von Grenzen voraus, die angesichts der Kontinuität der Natur („Natura non saltat." Die Natur macht keine Sprünge.) ad absurdum geführt werden kann. Die ist ein universelles, unvermeidbares und ohne Einsatz von Willkür unlösbares Problem!

Zur „Praxisnähe": Die Zuordnung zu den als zweckmäßig festgelegten Kategorien sollte unter Praxisbedingungen möglich sein.

Ohne Klassifizierung kommen wir nicht aus. Ohne explizite oder implizite Kategorisierung ist es nicht möglich

- Erfahrungen zu sammeln
- sinnvoll zu kommunizieren
- sinnvoll tätig zu werden
- zu leben (Jedes Lebewesen muss die Phänomene seiner Umwelt differenzieren, d.h., einer von mindestens zwei verschiedenen Kategorien [Beispiele: potentielle Nahrung oder potentieller Fressfeind, potentieller Geschlechtspartner oder potentieller Rivale] zuordnen; die Fähigkeit zur Bildung von Kategorien ist angeboren, nicht aber notwendigerweise die Inhalte.)

2.3 Testtheorie

Unter „Tests“ sollen hier alle Untersuchungen verstanden werden, die im Rahmen von Diagnostik durchgeführt werden, also auch die Prüfung auf das Vorliegen von klinischen Symptomen Der Begriff „Symptom“ wird in diesem Skript in sehr weit gefasster Bedeutung als diagnostisch verwertbares Attribut von kranken Individuen verwendet. Im angloamerikanischen Sprachraum werden „symptoms“ und „signs“ unterschieden. Erstere sind subjektive Phänomene („Beschwerden“), letztere objektiv oder zumindest intersubjektiv überprüfbare Attribute („Befunde“).

Eine Anmerkung vorweg: die folgenden Ausführungen zur Testtheorie mögen für Leute, die eine gewisse Freude an Berechnungen haben, interessant oder vielleicht sogar unterhaltsam sein. Es gibt aber ein Problem, das darin besteht, dass die zu diesen Berechnungen benötigten Informationen in der Realität selten vollständig zur Verfügung stehen. Es schadet aber nicht, die Prinzipien verstanden zu haben.

“It’s nice to know things.” Bertrand Russell

Tests sollen die Diagnose, also die Zuordnung eines konkreten Erkrankungsfalles zu einer nosologischen Einheit, also zu einer bestimmten Krankheit (K) erleichtern oder sicherer machen.

Im einfachsten Fall gibt es dabei folgende vier Möglichkeiten: Die Krankheit K ist entweder vorhanden oder nicht vorhanden. Die Teilmenge der Population, für die letzteres zutrifft, also "Krankheit nicht vorhanden", ist nicht identisch mit der Gruppe der Gesunden, schließt sie aber ein. Sie beinhaltet darüber hinaus jedoch auch alle Individuen mit anderen als der gesuchten Krankheit.

Wiederum im einfachsten Fall kann ein Test positiv oder negativ ausfallen (in einem solchen Fall spricht man von einem Test mit dichotomem Ergebnis). Durch diese doppelte Zweiteilung ergibt sich hinsichtlich der gesamten Population eine Vierteilung, die in einer Vierfeldertafel dargestellt werden kann(s. Tab. 2.3-1).

Tab. 2.3-1: Vierfeldertafel zur Darstellung des Ausfalls eines Tests bei Individuen mit einer bestimmten Krankheit K und bei solchen ohne sie

	Krankheit K vorhanden	**Krankheit K nicht vorhanden**
Testergebnis positiv	richtig positiv	falsch positiv
Testergebnis negativ	falsch negativ	richtig negativ

Es hat sich inzwischen weitgehend durchgesetzt, diese Vierfeldertafel auf diese Weise darzustellen, also Krankheitsstatus über den Spalten und Testergebnisse vor den Zeilen. Es gibt aber auch andere Darstellungen.

Auf diese Vierfeldertafel wird im weiteren Verlauf noch mehrmals zurückgegriffen werden. Es wird mitunter behauptet, Kliniker würden nur die linke Hälfte der Tabelle sehen, Epidemiologen auch die rechte.

Ist Krankheit K vorhanden, und der Test fällt positiv aus, spricht man von einem richtig positiven Ergebnis. Ist Krankheit K vorhanden, aber der Test fällt negativ aus, spricht man von einem falsch negativen Ergebnis. Ist Krankheit K nicht vorhanden (die Individuen also gesund oder an anderen Krankheiten leidend), aber der Test fällt positiv aus, spricht man von einem falsch positiven Ergebnis. Ist Krankheit K nicht vorhanden, und der Test fällt negativ aus, spricht man von einem richtig negativen Ergebnis.

Wie gut sich ein bestimmter Test zur Feststellung oder zum Ausschluss einer bestimmten Krankheit eignet, lässt sich anhand der im Folgenden zu besprechenden **Qualitätsmerkmale** quantifizieren. Diese Qualitätsmerkmale können zwar nur anhand der Untersuchung von geeigneten Gruppen (Populationen) ermittelt werden. Hier geht es jedoch zunächst um die Diagnostik am Individuum. Aspekte der Diagnostik an Populationen werden später besprochen.

2.3.1 Diagnostische Sensitivität

Die diagnostische Sensitivität eines bestimmten Tests im Hinblick auf eine bestimmte Krankheit entspricht dem Anteil der Individuen mit der gesuchten Krankheit, der im fraglichen Test positiv reagiert.

Tab. 2.3-2: Vierfeldertafel zur Darstellung des Ausfalls eines Tests bei Individuen mit einer bestimmten Krankheit K und bei solchen ohne sie

	Krankheit K vorhanden	**Krankheit K nicht vorhanden**
Testergebnis positiv	a (RP)	b (FP)
Testergebnis negativ	c (FN)	d (RN)
	a + c	b + d

Projiziert auf ein Individuum bedeutet die diagnostische Sensitivität die Wahrscheinlichkeit, dass ein zufällig ausgewähltes Individuum mit der gesuchten Krankheit in dem fraglichen Test positiv reagiert. Ob es korrekt ist, die an einer Population ermittelte Größe auf jedes Individuum dieser Population zu projizieren, ist jedoch nicht unumstritten. Wenn Interventionen an Individuen von Graden an Wahrscheinlichkeit abhängig gemacht werden sollen, ist man jedoch auf derartige Daten angewiesen.

Aus diesen Definitionen und Tabelle 2.3-2 ergibt sich die Formel:

$$\textbf{Diagnostische Sensitivität} = a/(a+c) \qquad (1)$$

Es handelt sich hier also um einen Dezimalbruch, der maximal den Wert 1 annehmen kann, nämlich dann, wenn das Feld „c“ nicht besetzt ist, es also keine falsch negativen Ergebnisse gibt. (In manchen Publikationen wird statt eines Dezimalbruchs eine Prozentangabe verwendet. Im Hinblick auf weitere Berechnungen ist ein Dezimalbruch aber deutlich besser geeignet.) Ein Test besitzt im Hinblick auf Krankheit K perfekte Sensitivität, wenn alle Individuen mit Krankheit K im Test positiv reagieren, also von diesem Test „erfasst“ werden.

2.3.2 Diagnostische Spezifität

Die diagnostische Spezifität eines bestimmten Tests im Hinblick auf eine bestimmte Krankheit entspricht dem Anteil der Individuen ohne die gesuchte Krankheit, der in dem fraglichen Test negativ reagiert. Diese Definition mag etwas merkwürdig anmuten, weil sie nicht mit der üblichen Vorstellung von einem spezifischen Test übereinzustimmen scheint. Dass das doch der Fall ist, geht hoffentlich aus dem Folgenden hervor.

Projiziert auf ein Individuum bedeutet die diagnostische Spezifität die Wahrscheinlichkeit, dass ein zufällig ausgewähltes Individuum ohne die gesuchte Krankheit in dem fraglichen Test negativ reagiert.

Aus diesen Definitionen und Tabelle 2.3-2 ergibt sich die Formel:

Diagnostische Spezifität = d/(b+d) **(2)**

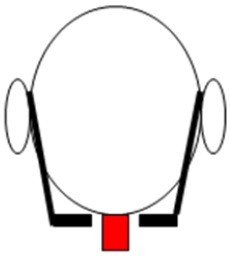
Auch die diagnostische Spezifität ist ein Dezimalbruch, der maximal den Wert 1 annehmen kann, nämlich dann, wenn das Feld „b“ nicht besetzt ist, es also keine falsch positiven Ergebnisse gibt. Ein Test besitzt im Hinblick auf Krankheit K perfekte diagnostische Spezifität, wenn nur Individuen mit der Krankheit K im Test positiv reagieren. Umgekehrt bedeutet das, dass ein positives Ergebnis mit Sicherheit auf das Vorliegen der fraglichen Krankheit deutet.

Diagnostische Sensitivität und Spezifität beziehen sich nur auf den Zusammenhang zwischen den Ergebnissen eines bestimmten Tests und dem Vorliegen einer bestimmten Krankheit.

Ein Test kann für eine Krankheit sehr sensitiv, für eine andere dagegen sehr wenig sensitiv sein. Oft sind sehr sensitive Tests wenig spezifisch. Ein Beispiel: Fast alle Kühe mit akuter Fremdkörperperitonitis haben Fieber, aber Fieber tritt bei sehr vielen anderen Krankheiten auf.

Frage 2.3.2-1: Im Allgemeinen werden diagnostische Sensitivität und Spezifität als konstante Eigenschaften eines Tests (im Hinblick auf eine bestimmte Krankheit) angesehen. Ist das gerechtfertigt?

Diagnostische Sensitivität und Spezifität eines Tests im Hinblick auf eine bestimmte Krankheit können nur ermittelt werden, wenn der wahre Status der Probanden im Hinblick auf die fragliche Krankheit bekannt ist.



Krankheit K vorhanden	Krankheit K nicht vorhanden
K	$\overline{K}$

Abb. 2.3-1: Rudolf, der Kliniker mit der roten Nase, sieht nur Individuen mit und solche ohne Krankheit K.

Das bringt bestimmte Probleme mit sich. Noch relativ harmlos ist das bei der Bestimmung der diagnostischen Sensitivität, da hier lediglich eine repräsentative Stichprobe von Individuen mit der bestimmten Krankheit untersucht werden muss. Repräsentativ ist eine Stichprobe, wenn jedes Individuum der Population die gleiche Chance hat, in der Stichprobe erfasst zu werden. Repräsentativität hat also nichts mit der Größe der Stichprobe zu tun, sondern bestimmt vielmehr den Grad an Sicherheit, mit der man die Population hinsichtlich Lage der zentralen Größe (Mittelwert) und der Streuung eines Parameters beschreiben kann.

Frage 2.3.2-2: Welche Auswirkung hat es, wenn zur Ermittlung der diagnostischen Sensitivität und Spezifität eines Tests im Hinblick auf eine bestimmte Krankheit K hundert Individuen mit K und hundert gesunde Freiwillige herangezogen werden?

Diagnostische Sensitivität und Spezifität werden unter dem Begriff Validität eines Tests zusammengefasst. Für den Anteil richtiger Ergebnisse an allen Ergebnissen gibt es verschiedene Ausdrücke, z.B. Richtigkeit oder Effektivität. Die nachstehende Formel bezieht sich auf die Darstellung in Tab. 2.3-2:

$$\textbf{Validität} = (a+d)/(a+b+c+d) \qquad (3)$$

Im täglichen Leben gibt es viele Relationen, die unter dem Aspekt von Sensitivität und Spezifität kritisch betrachtet werden können. Dabei zeigt sich, dass gelegentlich Fehler gemacht werden. Wenn zum Beispiel eine Firma den Werbeslogan „Du erkennst, dass ein Produkt von uns ist, wenn es gut ist“ hat, bedeutet das unter formalen Gesichtspunkten, dass das Attribut „gut“ perfekte Spezifität für die Produkte dieser Firma hat, also nur bei ihnen zu finden ist. Das hieße, dass die Produkte keiner anderen Firma „gut“ sein können, wäre aber auch vereinbar mit dem Umstand, dass fast alle Produkte dieser Firma nicht „gut“ sind. Ein anderer Werbeslogan („Alles von H. ist gut.“) setzt dagegen auf perfekte Sensitivität des Attributs „gut“ für die Produkte der Firma H. und lässt die Möglichkeit offen, dass auch

Produkte anderer Firmen gut sind. Das Attribut „gut“ ist also nicht spezifisch für die Produkte der Firma H.

2.3.3 Prädiktiver Wert (Vorhersagewert)

Obwohl Diagnostik, auch Labordiagnostik, schon seit lang er Zeit betrieben wird, wurde das zentral wichtige Konzept des Prädiktiven Werts anscheinend erst 1966 von VECCHIO beschrieben. Der prädiktive Wert des positiven Ausfalls eines bestimmten Tests (PWT+) im Hinblick auf eine bestimmte Krankheit K entspricht dem Anteil der Individuen mit positivem Testausfall, der auch tatsächlich Krankheit K hat.

Projiziert auf ein Individuum gibt der prädiktive Wert des positiven Testausfalls die Wahrscheinlichkeit an, dass ein positives Testergebnis tatsächlich das Vorliegen der Krankheit K anzeigt.

Tab. 2.3-3: Vierfeldertafel zur Darstellung des Ausfalls eines Tests bei Individuen mit einer bestimmten Krankheit und bei solchen ohne sie

	Krankheit K vorhanden	**Krankheit K nicht vorhanden**	
Testergebnis positiv	a (RP)	b (FP)	a+b
Testergebnis negativ	c (FN)	d (RN)	c+d
	a+c	b+d	

Aus dieser Definition und Tabelle 2.3-3 ergibt sich die Formel:

$$\textbf{Prädiktiver Wert des positiven Testausfalls} = a/(a + b) \qquad (4)$$

Auch der prädiktive Wert ist ein Dezimalbruch, der maximal der Wert 1 annehmen kann, und zwar wie die diagnostische Spezifität genau dann, wenn es keine falsch positiven Ergebnisse gibt, Feld „b“ also nicht besetzt ist.

Der prädiktive Wert des positiven Testausfalls liefert die Antwort auf die Frage: Wie sicher deutet ein positives Testergebnis auf das Vorliegen der gesuchten oder vermuteten Krankheit deutet?

Entsprechend ist der **prädiktive Wert des negativen Ausfalls** eines bestimmten Tests im Hinblick auf eine bestimmte Krankheit K der Anteil der Individuen mit negativem Testausfall, der tatsächlich frei von der fraglichen Krankheit ist. Mit anderen Worten: der prädiktive Wert des negativen Ausfalls eines bestimmten Tests im Hinblick auf eine bestimmte Krankheit ist die Wahrscheinlichkeit oder Sicherheit, dass ein negatives Testergebnis auf die Freiheit von der fraglichen Krankheit weist.

Daraus ergibt sich die Formel:

$$\textbf{Prädiktiver Wert des negativen Testausfalls} = d/(c + d) \qquad (5)$$

Während bei der Diskussion der diagnostischen Sensitivität und Spezifität sozusagen von oben in die Vierfeldertafel geschaut wird, nämlich in die Spalten „Krankheit K vorhanden“

und „Krankheit K nicht vorhanden“ (s. Abb. 2.3-1), wird bei der Diskussion des prädiktiven Wertes sozusagen von links in die Vierfeldertafel geschaut, nämlich in die beiden Zeilen „Test positiv“ und „Test negativ“ (s. Abb. 2.3-2). Diese Sicht entspricht der realen klinischen Situation, in der zwar das Testergebnis, nicht aber der wahre Krankheitsstatus eines Individuums bekannt ist. Denn sonst wäre eine weitere Diagnostik unnötig.

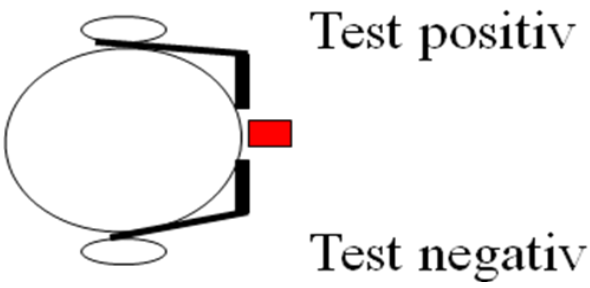


Abb. 2.3-2: Rudolf, der Kliniker mit der roten Nase, sieht nur Individuen mit positivem und solche mit negativem Testergebnis.

Dass der prädiktive Wert eines Testergebnisses (PWT+) im Hinblick auf eine bestimmte Krankheit von der diagnostischen Sensitivität (dSe) und Spezifität (dSp) des Tests abhängt, ist recht einleuchtend.

Zunächst zur Abhängigkeit von der dSe (s. Abb. 2.3-3).

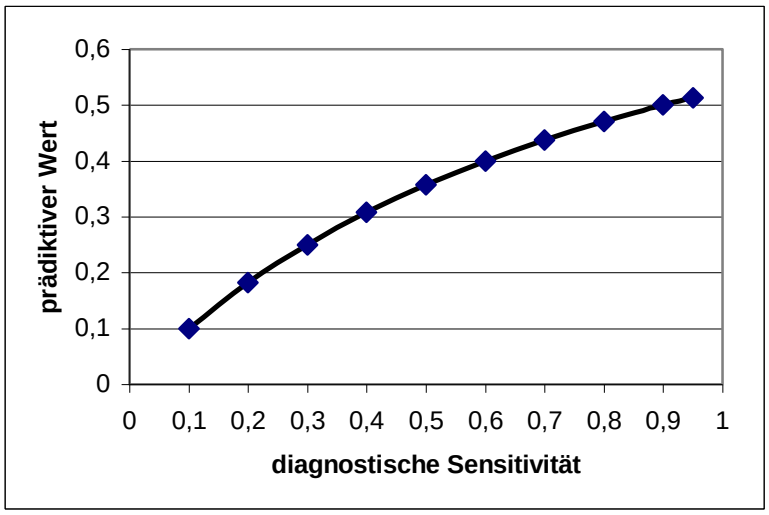


Abb. 2.3-3: Abhängigkeit des prädiktiven Wertes eines Tests von der diagnostischen Sensitivität (Prävalenz [siehe unten] = 0,1, dSp = 0,9)

Der prädiktive Wert ist also positiv mit der diagnostischen Sensitivität korreliert. Dasselbe gilt für sein Verhältnis zur dSp (s. Abb. 2.3-4).

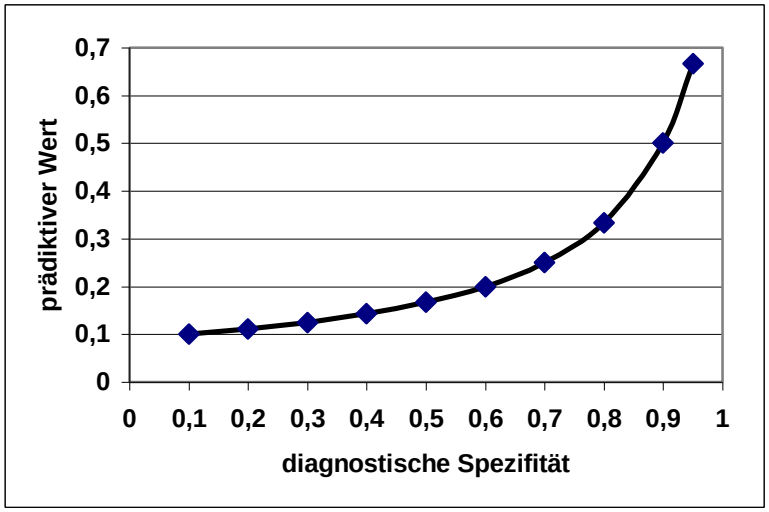


Abb. 2.3-4: Abhängigkeit des prädiktiven Werts eines Tests von der diagnostischen Spezifität (Prävalenz = 0,1; dSe = 0,9)

Weniger einleuchtend ist die Tatsache, dass der prädiktive Wert auch von der Prävalenz abhängt, also von der momentanen Häufigkeit der fraglichen Krankheit in der Population, oder von ihrer so genannten Apriori-Wahrscheinlichkeit oder Vor-Test-Wahrscheinlichkeit (s. Abb. 2.3-5). Dabei zeigt sich auch, dass die dSp den PWT+ stärker beeinflusst als die dSe.

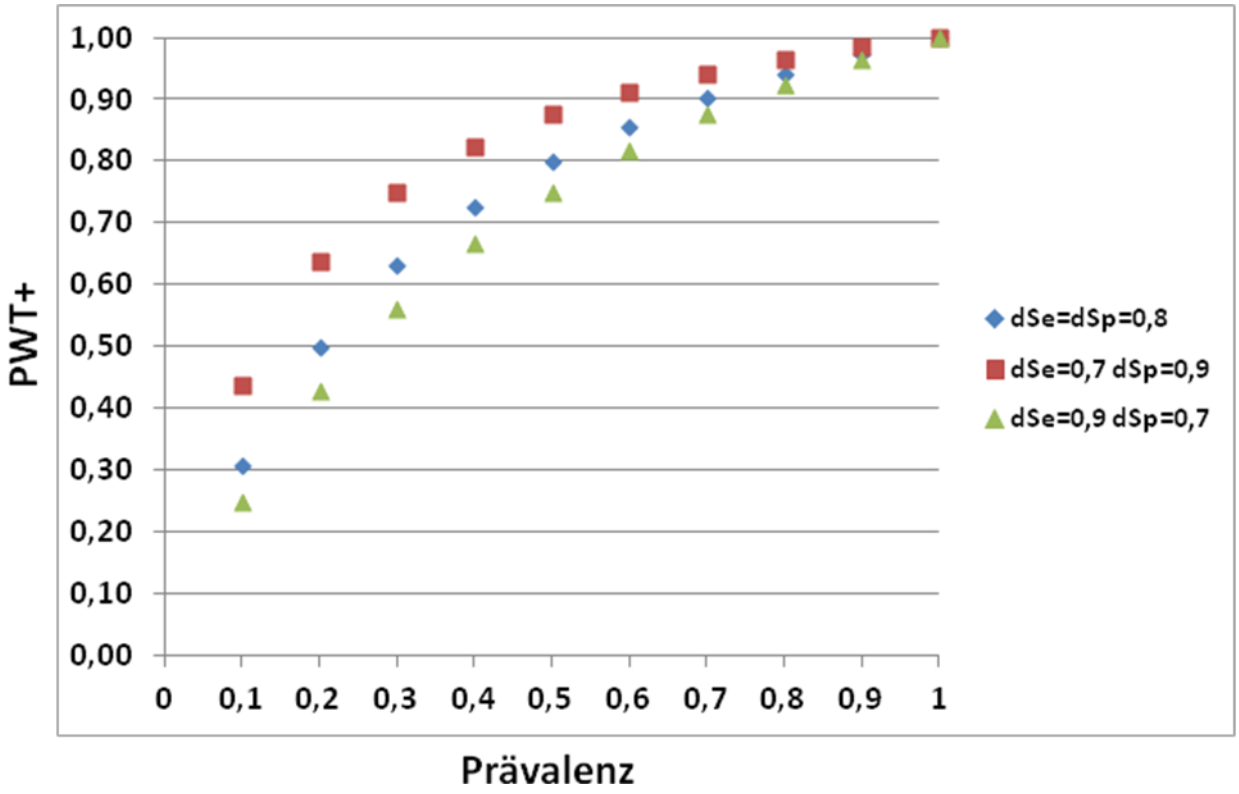


Abb. 2.3-5: Abhängigkeit des prädiktiven Werts des positiven Testergebnisses von der Prävalenz der fraglichen Krankheit bei verschiedenen Werten für dSe und dSp

Zur Illustration ein Beispiel: Ein klinischer Forscher entwickelt einen neuen Test für die Krankheit K. Um die Validität (Aussagekraft) des Tests zu untersuchen, führt er ihn bei 100 Individuen mit der fraglichen Krankheit und bei 100 gesunden Kontrollpersonen durch (Freiwillige, Studenten, Krankenhauspersonal; s. Tab. 2.3-4).

Tab. 2.3-4: Ausfall eines Test bei Individuen mit einer bestimmten Krankheit und bei einer gesunden Kontrollgruppe

	Krankheit K vorhanden	**Gesunde Kontrollgruppe**	
Testergebnis positiv	95	2	97
Testergebnis negativ	5	98	103
	100	100	200

Wie hier dargestellt, fällt der Test bei 95 % der Kranken positiv (dSe = 95/100 = 0,95) und bei 98 % der Gesunden negativ aus (dSp = 98/100 = 0,98).

Der Erfinder stellt erfreut fest, dass ein positives Ergebnis mit 98 %iger, also fast vollkommener Sicherheit auf das Vorliegen der Krankheit weist (PWT+ = 95/97 = 0,98). Stolz publiziert unser forscher Forscher seine Ergebnisse. Ein Landpraktiker ist bei der Lektüre begeistert, denn nun, so glaubt er, hat er endlich einen guten Test zum Nachweis dieser sonst so schwer zu diagnostizierenden Krankheit K. Er wendet ihn bei seinen Patienten an, unter denen die Prävalenz dieser Krankheit mit etwa 1 % (0,01) angenommen wird (s. Tab. 2.3-5).

Tab. 2.3-5: Ausfall eines Tests bei Individuen mit einer bestimmten Krankheit und bei solchen ohne sie

	Krankheit K vorhanden	**Krankheit K nicht vorhanden**	
Testergebnis positiv	95	198	293
Testergebnis negativ	5	9702	9707
	100	9900	10000

Hier jedoch weist ein positives Testergebnis nur noch in einem Drittel der Fälle auf das Vorliegen der Krankheit hin (PWT+ = 95/293 = 0,32), obwohl Sensitivität und Spezifität des Tests unverändert sind. Mit anderen Worten: bei zwei von drei Patienten würde ein positives Testergebnis zu einer Fehldiagnose führen. In der Realität wäre PWT+ noch viel geringer, weil die „Kontrollgruppe“ ja aus Patienten bestünde, also Individuen mit anderen, differentialdiagnostisch relevanten Krankheiten, bei denen der Test möglicherweise ebenfalls positiv reagiert. Dadurch würde der Anteil falsch positiver Ergebnisse, also die Zahl in Feld b, erhöht.
Die klinische Untersuchung kann formal auch unter dem Gesichtspunkt betrachtet werden, dass mit ihr meist einfach und rasch die Wahrscheinlichkeit für eine bestimmte Krankheit K vor der Durchführung aufwändiger Diagnostik-Verfahren und damit deren prädiktiver Wert erhöht werden kann, indem Individuen, die sicher nicht an „K“ leiden, „ausgesondert“ werden. Damit werden unter den verbleibenden Patienten diejenigen mit K „angereichert“.

Tab. 2.3-6: Ausfalls eines Tests bei Individuen mit einer bestimmten Krankheit und solchen ohne sie

	Krankheit K vorhanden	**Krankheit K nicht vorhanden**	
Testergebnis positiv	Pr*dSe	(1-Pr)*(1-dSp)	
Testergebnis negativ	Pr*(1-dSe)	(1-Pr)*dSp	
	Pr	1-Pr	

Mit Hilfe der drei jeweils als Dezimalbruch angegebenen Größen Prävalenz (Pr), dSe und dSp lässt sich eine allgemeine Formel für PWT+ aufstellen (s. Tab. 2.3-6).

PWT+ = (Pr*dSe)/[(Pr*dSe) + (1-Pr)*(1-dSp)] **(6)**

Erklärung: Es handelt sich um eine andere Darstellung der oben genannten Definition des PWT+ [a/(a+b)]. Der Ausdruck Pr*dSe entspricht dem Feld a (richtig positiv), denn er stellt den Anteil der Individuen mit der Krankheit (Prävalenz Pr = (a+c)/(a+b+c+d) = (RP + FN)/Gesamtzahl) dar, der im Test positiv reagiert. (1-Pr) ist der Rest der Population, also alle Individuen, die nicht die fragliche Krankheit (aber möglicherweise andere Krankheiten!) haben. Da dSp multipliziert mit dem Anteil der nicht an der fraglichen Krankheit erkrankten (1-Pr) den im Test negativ reagierenden Anteil dieser Individuen angibt (was dem Feld d oder „richtig negativ“ entspricht), muss (1-Pr) mit (1-dSp) multipliziert werden, also dem Komplement von dSp. Das Produkt (1-Pr)*(1-dSp) entspricht dann dem Feld b.

In der Realität liegen nur die Befunde „Testergebnis positiv“ und „Testergebnis negativ“ vor, wobei „Testergebnis positiv“ lediglich die scheinbare Prävalenz (sPr) darstellt. Anhand der Angaben in Tabelle 2.3-6 lässt sich die wahre Prävalenz (Pr) ausrechnen:

sPr = Pr*dSe + (1-Pr)*(1-dSp) **(7)**
sPr = Pr*dSe + 1- Pr – dSp +Pr*dSp
sPr = Pr*(dSe + dSp -1) + 1 – dSp
Pr = (sPr + dSp -1)/(dSe + dSp -1) **(7a)**

Ein Beispiel mit den Zahlen aus Tabelle 2.3-5:
Pr = (0,0293 + 0,98 -1)/(0,95 + 0,98 -1)
Pr = 0,01

Frage 2.3.3-1: Nachdem nun die wichtigsten Testqualitätskriterien vorgestellt wurden, kann man sich fragen, wann welches Kriterium wichtig ist. Wann braucht man einen Test mit hoher diagnostischer Sensitivität, wann einen mit hoher diagnostischer Spezifität?

Frage 2.3.3-2: Wie hoch ist der prädiktive Wert des Münzwurfs, also des „Tests“, der darin besteht, in einer unklaren klinischen Situation eine Münze zu werfen und „Zahl“ als Hinweis für das Vorliegen der Krankheit „Mysteriose“ zu werten?

Bisher wurde stets von ***diagnostischer*** Sensitivität und Spezifität gesprochen (oder vielmehr geschrieben). Die Begriffe Sensitivität und Spezifität gibt es auch noch in einem anderen Zusammenhang, nämlich in der Labordiagnostik (siehe Skriptum „Klinische Pathophysiologie und Labordiagnostik“).

Die ebenfalls in der Labordiagnostik wichtigen Begriffe der Präzision und Richtigkeit werden ebenfalls im Skriptum „Klinische Pathophysiologie und Labordiagnostik besprochen.

2.3.4 Likelihood ratio

Eine andere Möglichkeit, die Validität eines Tests zu quantifizieren, ist die so genannte Likelihood ratio (Lr). Sie bezeichnet das Verhältnis der Wahrscheinlichkeit eines positiven (oder negativen) Testausfalls bei Individuen mit der fraglichen Krankheit zu der Wahrscheinlichkeit eines positiven (oder negativen) Testausfalls bei Individuen ohne die fragliche Krankheit. Aus dieser Definition ergibt sich die Formel

Lr = [a/(a+c)]/[b/(b+d)] **(8)**

Die Lr ist – im Gegensatz zum PW – nicht von der Prävalenz abhängig. Ein Manko, das sowohl dem Konzept des PW als auch dem der Lr anhaftet, besteht darin, dass die klinisch relevante dSp (im Gegensatz zu der an gesunden Individuen ermittelte) so gut wie nie bekannt ist, unter anderem deshalb, weil sie von der jeweiligen klinischen Situation (mit unterschiedlichen mehr oder weniger wahrscheinlichen Differentialdiagnosen) abhängt. Auch wenn diese Einschränkung der Nützlichkeit dieser Konzepte ernüchternd klingen mag, sollte man darüber die Tatsache nicht aus den Augen verlieren, dass Beachtung und Anwendung dieser Konzepte zum korrekten Umgang mit Unsicherheit anhalten.

2.3.5 Kombination von Tests

Bisher wurde von der vereinfachten Situation ausgegangen, dass zur Diagnostik einer Krankheit nur jeweils ein Test herangezogen wird. Das ist ziemlich unrealistisch. Daher soll nun auch die Möglichkeit der Kombination von Tests besprochen werden.

Zunächst zwei Grafiken, die in sehr optimistischer Weise den Nutzen von Testkombinationen zeigen. In Abb. 2.3-6 ist dargestellt, wie durch die Kombination mehrerer Tests (T1 bis Tn) die Sicherheit über das Vorliegen einer bestimmten Krankheit K schrittweise gesteigert werden kann. Ausgangswert ist die Prävalenz oder Apriori-Wahrscheinlichkeit p(K), also die Wahrscheinlichkeit der fraglichen Krankheit vor der Durchführung des ersten Tests. Das Ende ist erreicht, wenn die gewünschte Schwelle von Wahrscheinlichkeit = Sicherheit erreicht oder überschritten wird. Die Inkremente an Wahrscheinlichkeit für das Vorliegen von Krankheit K nach positivem Ausfall von Test 1 bzw. Test 2 werden durch a bzw. b dargestellt.
Die Darstellung ist insofern idealisiert, als falsch negative Ergebnisse nicht berücksichtigt werden.

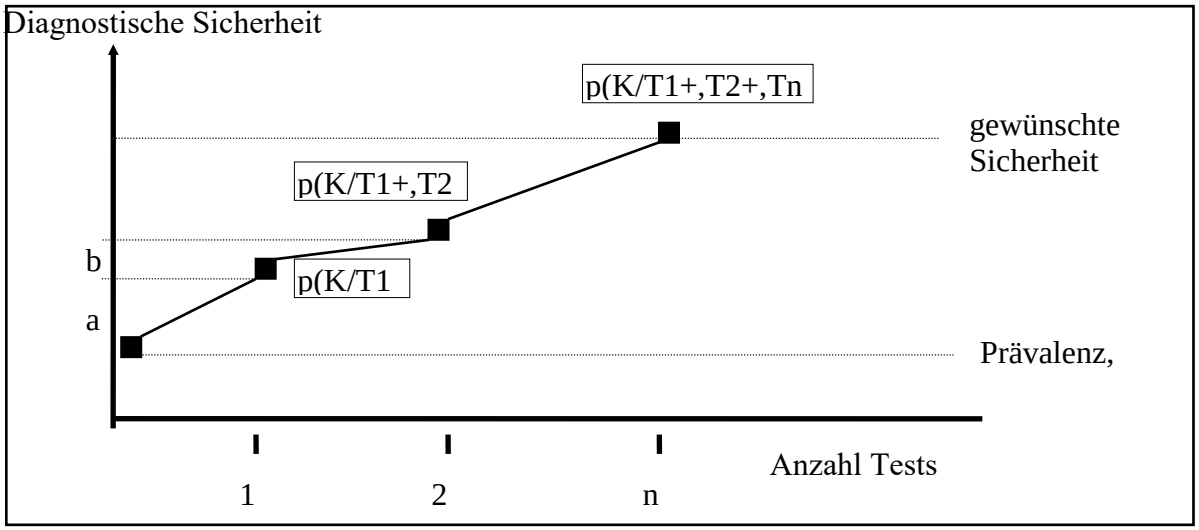


Abb. 2.3-6: Serielle Kombination von Tests zur schrittweisen Steigerung der diagnostischen Sicherheit

Die Apriori-Wahrscheinlichkeit ist nicht fix, sondern hängt unter anderem vom Ort (Landpraxis oder Spezialklinik) und vom Zeitpunkt (Ausbruch der fraglichen Krankheit, endemische Situation oder anerkannte Freiheit des Gebietes von der Krankheit) ab. Wenn in eine Klinik für Wiederkäuer eine Kuh eingeliefert wird, entspricht die Apriori-Wahrscheinlichkeit, dass sie Labmagenverlagerung hat (LMV) hat, dem Anteil von Kühen mit LMV an der Gesamtzahl der in einem bestimmten Zeitraum eingelieferten Kühe. In die subjektive Einschätzung der Wahrscheinlichkeit einer LMV durch die untersuchende Tierärztin gehen aber auch ohne Berechnungen fast unbewusst sofort weitere Faktoren ein, so z.B. Rasse, Laktationsstadium, Angaben aus dem Vorbericht.
Es zeigt sich, dass das Konzept der Apriori-Wahrscheinlichkeit (Prävalenz) theoretisch leicht verständlich, im klinischen Alltag aber nicht leicht anwendbar ist. Ist sie aber erst einmal (auf welche Weise auch immer – z.B. im Rahmen von Dissertationsprojekten, durchgeführt durch die klinisch Unerfahrensten) ermittelt, können die Epidemiologen damit wunderbare Berechnungen durchführen.

Die zweite Grafik (Abb. 2.3-7) stellt dar, dass es unter bestimmten Umständen möglich sein kann, durch Anwendung einer Testkombination, hier Test 1 und Test 2, zwei Gruppen, hier K1 und K 2, voneinander zu trennen, was durch alleinige Anwendung eines der beiden Tests nicht möglich wäre. Die auf die beiden "Testdimensionen", also Koordinaten, projizierten Ergebnisse für K1 und K2 überschneiden sich im Bereich b-c (Test 1) und im Bereich f-g (Test 2). Mit Hilfe bestimmter Verfahren, z. B. der Diskriminanzanalyse, ist es jedoch möglich, eine dritte Dimension zu errechnen (Gerade G in der Abbildung), hinsichtlich welcher der Grad der Überschneidung mehr oder weniger reduziert ist, im günstigsten Fall auf Null, wie hier dargestellt (überschneidungsfreier Bereich i-j). Das bedeutet, dass die beiden Gruppen jetzt eindeutig voneinander getrennt werden können.

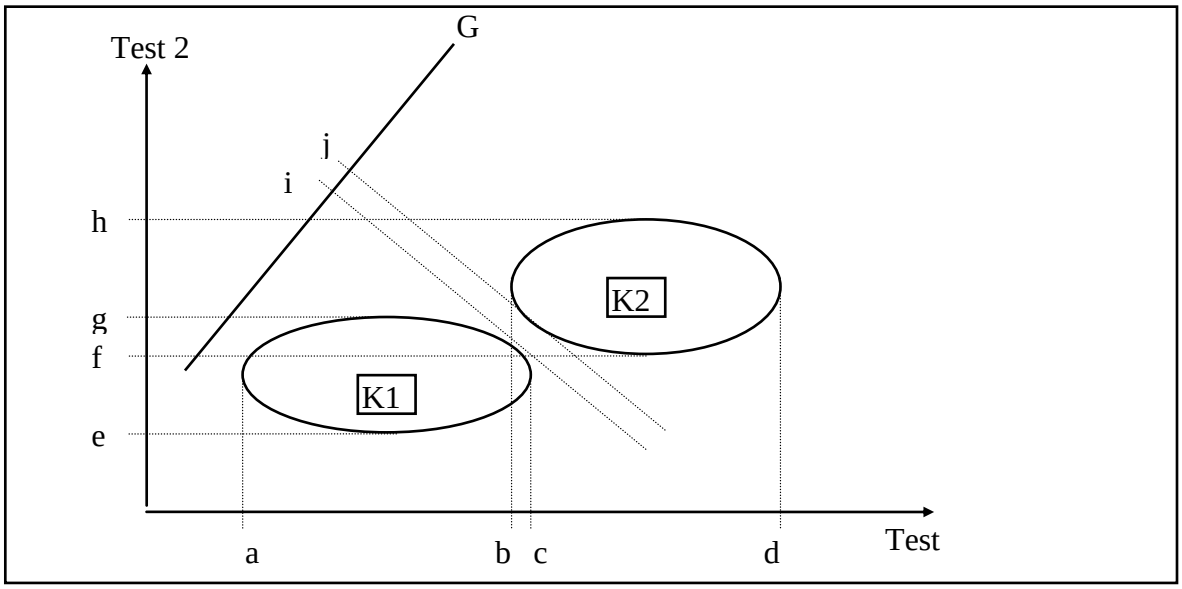


Abb. 2.3-7: Parallele Kombination von Tests zur besseren Trennung zweier Populationen

Zusätzliche Tests sind fast immer mit zusätzlichen Kosten und/oder Risiken verbunden, bringen jedoch unter Umständen nur wenig zusätzliche Information. Der Einfachheit halber wird die Problematik bei der Durchführung zweier Tests beschrieben.

Folgende Aspekte sind mit der Anwendung mehrerer (zweier) Tests verbunden und bereiten mitunter Probleme:

- Reihenfolge
- diskrepante Ergebnisse
- Korrelation

Zur Reihenfolge:
Zwei Tests A und B können gleichzeitig, also parallel, oder nacheinander, also seriell durchgeführt werden, wobei es im letzteren Fall zwei Möglichkeiten gibt, nämlich A-B oder B-A.

Vorteile paralleler Durchführung:
- Zeitersparnis
- eventuell Kostenersparnis (im Rahmen einer Testpalette)

Vorteile serieller Durchführung:
- Kostenersparnis (weil unter Umständen weniger Tests erforderlich sind)
- höhere diagnostische Effizienz

Es gibt auch eine Art vernünftiger Folge von Tests (s. Abb. 2.3-8).

sensitiv	einfach	billig	ungefährlich
⇓	⇓	⇓	⇓
spezifisch	aufwändig	teuer	gefährlich

Abb. 2.3-8: Vernünftige Abfolge von Tests

Zu diskrepanten Ergebnissen:
Wenn zwei Tests durchgeführt werden, gibt es bei der Interpretation der Ergebnisse keine Schwierigkeiten, wenn sie übereinstimmen, also beide entweder positiv oder negativ sind. Anders ist das, wenn die Ergebnisse nicht übereinstimmen, also diskrepant sind. Im Prinzip gibt es dann zwei Möglichkeiten des Vorgehens.

Tab. 2.3-7: Gesamtergebnis beim Durchführen von zwei Tests (disjunktiv)

	Test A positiv	**Test A negativ**
Test B positiv	Gesamtergebnis positiv	Gesamtergebnis positiv
Test B negativ	Gesamtergebnis positiv	Gesamtergebnis negativ

Bei dem in Tab. 2.3-7 dargestellten Vorgehen wird das Gesamtergebnis als positiv gewertet, wenn wenigstens ein Einzeltest ein positives Ergebnis liefert. Man spricht von einem *disjunktiven Positivitätskriterium* (englisches Stichwort: believe the positive!).

Bei der anderen Strategie (s. Tab. 2.3-8) wird das Gesamtergebnis nur dann als positiv gewertet, wenn beide Einzeltests positive Ergebnisse liefern. Man spricht von einem *konjunktiven Positivitätskriterium* (englisches Stichwort: believe the negative!).

Tab. 2.3-8: Gesamtergebnis beim Durchführen von zwei Tests (konjunktiv)

	Test A positiv	**Test A negativ**
Test B positiv	Gesamtergebnis positiv	Gesamtergebnis negativ
Test B negativ	Gesamtergebnis negativ	Gesamtergebnis negativ

In der folgenden Grafik (Abb. 2.3-9) bedeckt die Krankheit K 24 der insgesamt 100 Felder. Die Prävalenz oder Apriori-Wahrscheinlichkeit von Krankheit K beträgt also 24/100 = 0,24.

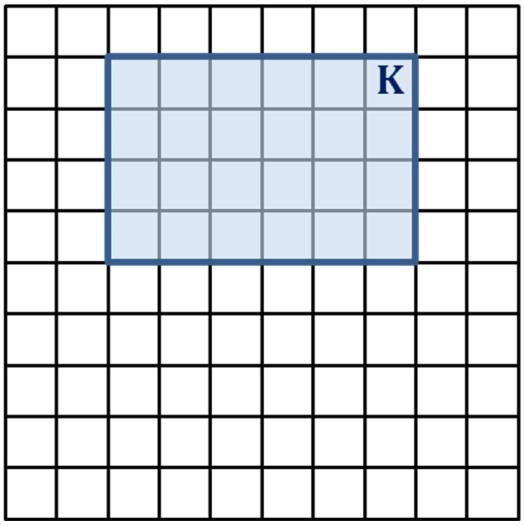


Abb. 2.3-9: Prävalenz von Krankheit K

In der Abb. 2.3.10 sind Individuen, welche Symptom S1 oder S2 oder beide zeigen, markiert.

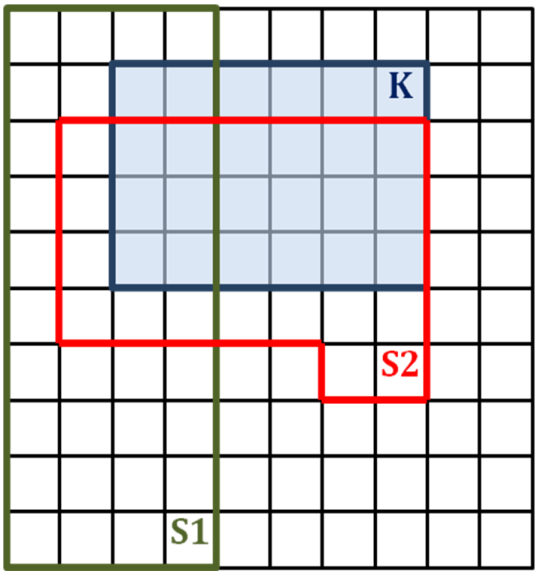


Abb. 2.3-10: Krankheit K, Symptom S1 und Symptom S2

Frage 2.3.4-1: Wie hoch sind diagnostische Sensitivität und Spezifität von Symptom S1 und S2 im Hinblick auf Krankheit K?

Die Sensitivitäten und Spezifitäten der beiden Einzelsymptome sind (nach Beantwortung von Frage 2.3.4-1) jetzt bekannt.

Frage 2.3.4-2: Wie sind diagnostische Sensitivität und Spezifität der Symptomenkombination bei disjunktivem und wie bei konjunktivem Positivitätskriterium?

In der nachstehenden Tabelle 2.3-9 sind die Wirkungen der beiden Positivitätskriterien auf diagnostische Sensitivität und Spezifität der Symptomenkombination zusammengefasst.

Tab. 2.3-9: Auswirkung des gewählten Positivitätskriteriums auf dSe und dSp

	Disjunktives Positivitätskriterium	**Konjunktives Positivitätskriterium**
dSe	steigt	sinkt
dSp	sinkt	steigt

Welches Positivitätskriterium das bessere ist, hängt von der Apriori-Wahrscheinlichkeit der in Frage stehenden Krankheit und von den Konsequenzen falsch positiver und falsch negativer Zuordnungen ab.

Frage 2.3.4-3: In welcher Situation ist ein disjunktives, in welcher ein konjunktives Positivitätskriterium vorzuziehen?

Zur Korrelation:
Im vorstehenden Abschnitt wurde die dSe der Symptomenkombination S1 und S2 bei Verwendung des disjunktiven Positivitätskriteriums „empirisch“ (durch Abzählen) ermittelt (s. Abb. 2.3-10). Man könnte sie nach Kenntnis der dSe der beiden Einzelsymptome auch aufgrund folgender Überlegung berechnen. Gemäß der dSe von S1 von 0,33 werden 33 % der 24 Individuen mit K (also 8) schon durch Prüfung auf S1 erfasst. Von den übrigen 67 % (also 16) der Individuen mit K erfasst die Prüfung auf S2 gemäß der dSe von S2 nochmals 75 % (also 12). Damit werden 20 der 24 Individuen mit K durch Prüfung auf S1 und/oder S2 erfasst. Theorie und „Realität“ stimmen also überein. Als Formel ausgedrückt:

$$\mathbf{dSe_{(S1,S2)} = dSe_{S1} + (1 - dSe_{S1})*dSe_{S2}} \qquad \mathbf{(9)}$$

In Abbildung 2.3-11 ist die dSe von S3 im Hinblick auf K 9/24 = 0,375. Wenn die dSe von der Kombination S1 und S3 unter Verwendung des disjunktiven Positivitätskriteriums in analoger Weise wie oben berechnet wird, ergibt sich 0,33 + (1 - 0,33)*0,375 = 0,58. Das würde 14 der 24 Individuen mit K entsprechen. Tatsächlich werden aber nur 11 und nicht 14 Individuen mit K erfasst, d.h., die dSe beträgt 11/24 = 0,46. Die dSp der beiden Symptome allein beträgt 32/76 = 0,42 für S1 und 11/76 = 0,14 für S3. Unter Verwendung von dPK beträgt die Spezifität der Kombination 42/76 = 0,55, ist also höher als die dSp der beiden einzelnen Symptome. Unter Verwendung des kPK beträgt die dSp der Kombination 9/76 = 0,12, ist also niedriger als die dSp der beiden einzelnen Symptome.

Offensichtlich liegen bei der Kombination S1 und S2 (Abb. 2.3-10) andere Verhältnisse vor als bei der Kombination S1 und S3. Die Wahrscheinlichkeiten für S1 und S2 in der Population sind 0,4 bzw. 0,3. Tatsächlich ist die Wahrscheinlichkeit, dass S1 und S2 vorliegen 0,12, entspricht also dem Produkt der Einzelwahrscheinlichkeiten. Auch unter den Individuen mit Krankheit K gilt diese Beziehung (1/3 * 3/4 = 6/24). Dies spricht dafür, dass die beiden Symptome S1 und S2 unabhängig voneinander vorkommen.

Die Wahrscheinlichkeit des gemeinsamen Auftretens voneinander unabhängiger Ereignisse ist gleich dem Produkt der Einzelwahrscheinlichkeiten.

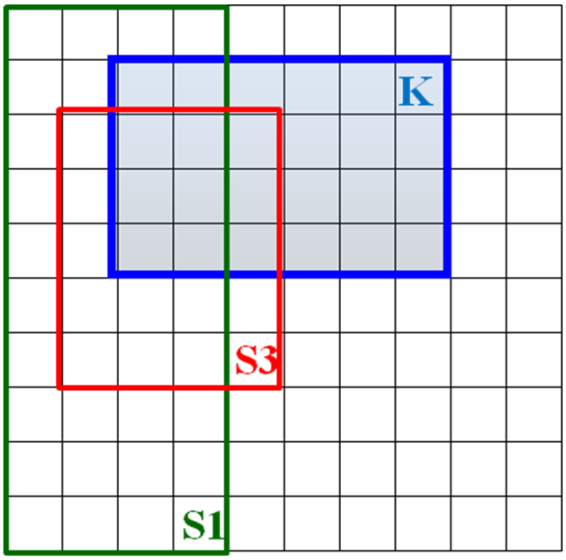


Abb. 2.3-11: Krankheit K, Symptom S1 und Symptom S3

Bei der Kombination S1 und S3 sieht es etwas anders aus. S3 hat in der Population eine relative Häufigkeit von 20/100 = 0,2. Bei Unabhängigkeit von S1 und S3 wäre ein Überlappungsbereich von 0,4*0,2 = 0,08 (also 8 Felder) zu erwarten. Tatsächlich zeigen aber 15 Individuen S1 und S3. Ein weiterer Aspekt der Unabhängigkeit wäre, dass die Wahrscheinlichkeit des einen Symptoms nicht durch das Vorliegen des anderen beeinflusst wird. Formal ausgedrückt, bedeutet das p(Sx/Sy) = p(Sx). Die Wahrscheinlichkeit für S1 beträgt 40/100 = 0,4. Die Wahrscheinlichkeit für S1 unter der Bedingung, dass S3 vorliegt, also p(S1/S3), beträgt aber 15/20 = 0,75 und ist damit deutlich höher als p(S1). S1 und S3 kommen also nicht unabhängig voneinander vor, sondern sind positiv miteinander korreliert. Negative Korrelation ist gekennzeichnet durch die Relation p(Sx/Sy) < p(Sx).

Die Frage, ob zwei Tests voneinander unabhängig sind, oder ihre Ergebnisse mehr oder weniger miteinander korrelieren, lässt sich nicht immer von vornherein beantworten, sondern muss empirisch ermittelt werden. Man kann davon ausgehen, dass Tests, die sehr ähnliche Funktionen eines pathophysiologischen Geschehens messen, mehr oder weniger korreliert sind.

In Tabelle 2.3-10 sind Ergebnisse von verschiedenen Verfahren zur Untersuchung auf Paratuberkulose einer größeren Zahl von Rindern. In den Randfeldern sind die Anteile positiver oder fraglicher Ergebnisse in den einzelnen Verfahren angegeben. In der Tabelle selbst sind Anteile der in einem oder beiden Verfahren einer Kombination positiv oder fraglich reagierenden Probanden sowie der unter Annahme der Unabhängigkeit zu erwartenden Ergebnisse dargestellt. Die Berechnung der theoretisch zu erwartenden Ergebnisse erfolgt nach Formel (9). Die Daten sind mit der Annahme vereinbar, dass die Tests in der Tat weitgehend unabhängig voneinander sind. Im Fall der beiden serologischen Tests (ELISA und KBR) ist das besonders überraschend.

Tab. 2.3-10: Proportionen von (tatsächlichen und theoretisch zu erwartenden) Kombinationen positiver und fraglicher Ergebnissen verschiedener Test für Paratuberkulose bei Rindern.

	Johnin (9,2 %)	KBR (4,8 %)	ELISA (15,2 %)
Kultur (1,8 %)	**10,6 %** **10,8 %**	**6,0 %** **6,5 %**	**16,0 %** **16,7 %**
Johnin (9,2 %)		**13,4 %** **13,6 %**	**22,3 %** **23,0 %**
KBR (4,8 %)			**16,6 %** **19,3 %**
Alle	**24,4 %** **28,0 %**		

Je größer die Korrelation der Ergebnisse, desto geringer die zusätzliche Information, die durch die Kombination gegenüber einem Einzeltest zu gewinnen ist. Der Informationsgewinn durch Anwendung eines weiteren Tests ist maximal, wenn die Tests auf voneinander unabhängige Größen abzielen.

Bei Anwendung des disjunktiven Positivitätskriteriums sinkt die Sensitivität bei positiver Korrelation und sie steigt bei negativer Korrelation. Wie immer, verhält es sich mit der Spezifität umgekehrt. Und noch einmal umgekehrt verhält es sich bei Anwendung des konjunktiven Positivitätskriteriums

2.3.6 Indikation für Tests

Man könnte geneigt sein zu glauben, dass in jeden Fall Diagnostik bis zur völligen Klärung der vorliegenden Krankheit (und gegebenenfalls noch weiter zur Abklärung der Prognose) indiziert ist. Ob das wirklich so ist, müsste anhand von kontrollierten Studien ermittelt werden, in denen Patienten nach dem Zufallsprinzip einer von zwei Gruppen zugeteilt werden, wobei eine dem zu prüfenden Test unterzogen wird, die andere aber nicht. Verglichen wird der Ausgang der Erkrankung.
Es ist leicht einzusehen, dass ein Test umso eher eingesetzt wird, je ungefährlicher er ist, je höher seine Validität ist, je wirksamer und ungefährlicher die Therapie ist, die bei positivem Testausfall durchgeführt wird und je schwerwiegender die Krankheit ohne Behandlung ist. Wenn dagegen die nach positivem Testergebnis indizierte Therapie mit einem erheblichen Risiko behaftet ist, kann es besser sein, auf einen Test zu verzichten.

2.4 Nosographie

Die Beschreibungen von Krankheiten in Lehrbüchern sind im Prinzip Angaben über die diagnostische Sensitivität von bestimmten Symptomen bei den jeweiligen Krankheiten. Hierbei gibt es wenigstens 5 verschiedene Möglichkeiten (s. Abb. 2.4-1).

Krankheiten	Symptome
A	a, b, c, ...
B	a, d, e, ...
A	a (regelmäßig), b (gelegentlich), c (selten)
B	a (gelegentlich), d (häufig), e (selten)
A	a (95 %), b (23 %), c (9 %)
B	a (17 %), d (73 %), e (5 %)
A	a (54/57), b (13/57), c (5/57)
B	a (11/63), b (46/63), e (3/63)
A	a (54/57), b (13/57), c (5/57), d (0/57)

Abb. 2.4-1: Möglichkeiten zur Beschreibung von Krankheiten

Bei der ersten werden die Symptome nur aufgezählt. Das würde bedeuten, dass diese Symptome in jedem Fall der Krankheit vorhanden sind, ihre dSe also 1,0 (oder 100 %) beträgt, was nicht wahrscheinlich ist. Außerdem besteht das Problem, dass *de facto* identische Krankheitserscheinungen mitunter von verschiedenen Autoren mehr oder weniger unterschiedlich beschrieben werden, was selten verwendeten Formulierungen eine scheinbar hohe Spezifität verleiht.

Bei der zweiten Möglichkeit werden grobquantitative Angaben zur Häufigkeit in Form von Ausdrücken wie "häufig", "gelegentlich" oder "selten" gemacht. Das Problem bei solchen Begriffen besteht darin, dass Menschen nachweislich sehr unterschiedliche Vorstellungen mit ihnen verbinden.

Bei der dritten Möglichkeit werden die Häufigkeiten der einzelnen Symptome in Prozentzahlen angegeben.

Bei der vierten Möglichkeit wird genau angegeben, wie viele von wie vielen untersuchten Individuen das betreffende Symptom aufwiesen.

Bei der fünften Möglichkeit werden auch Beobachtungen über Fehlen von Symptomen quantifiziert und so einer statistischen Bearbeitung zugänglich gemacht. Die letzten beiden sind die besten Verfahren. Denn erstens kann sich sowohl der Autor als auch der Leser einen Eindruck von der Basis machen, auf der diese Angaben beruhen, und zweitens können solche Angaben ohne weiteres mit denen anderer Autoren zu erweiterten Kasuistiken zusammengefasst werden.

Es gibt im Wesentlichen drei Möglichkeiten, Symptome darzustellen:

2.4.1 Symptome als konstante, qualitative Merkmale

Es wird so getan, als würden die Symptome schlagartig und dauerhaft auftreten, wie Glühbirnen, die angeschaltet werden (s. Abb. 2.4-2)

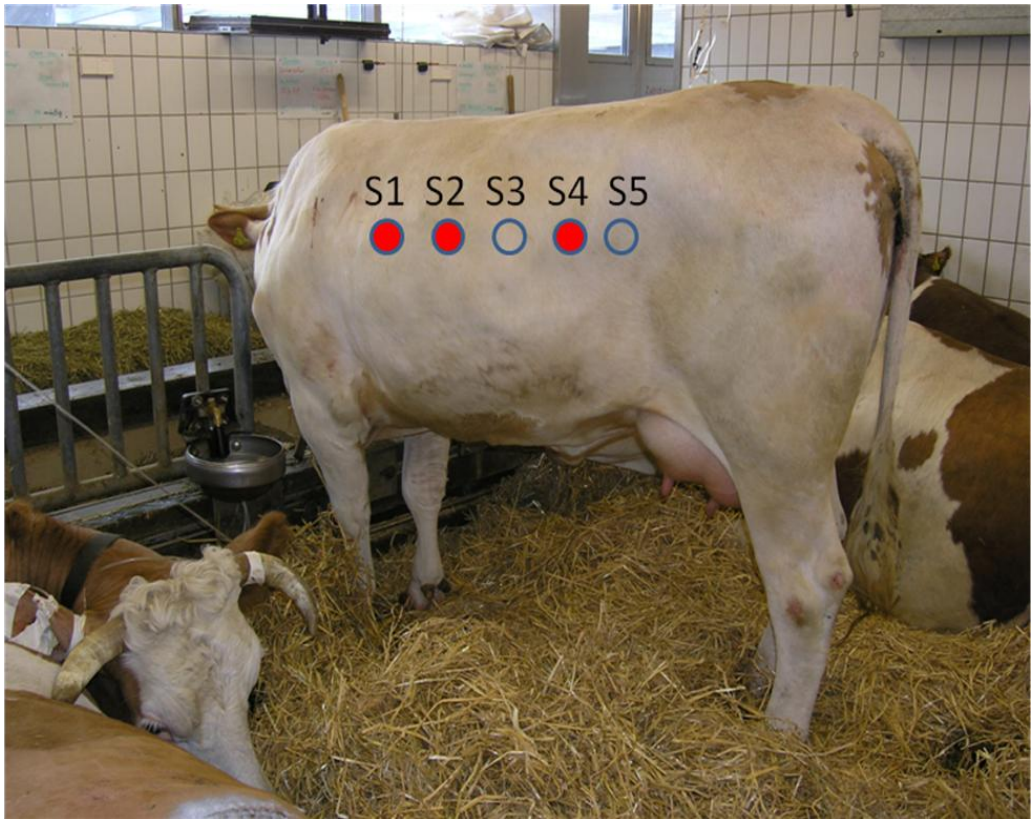


Abb. 2.4-2: Illustration konstanter, qualitativ erfasster Symptome (Erläuterung s. Tab. 2.4-2)

Die formalisierte Verarbeitung ist einfach. Auch sehr viele Symptome können leicht in Form der so genannten BOOLEschen Algebra bearbeitet werden:

Tab. 2.4-2: Diagnostik auf der Basis der BOOLEschen Algebra

	Krankheit K1	Krankheit K2	Krankheit K3	Krankheit K4
Symptom S1	1	0	1	1
Symptom S2	0	1	1	0
Symptom S3	1	1	0	0
Symptom S4	0	1	1	1
Symptom S5	1	1	0	1

Diese Art von Krankheiten-Symptomen-Matrix liegt Computer-Diagnostik-Programmen zu Grunde, die deterministisch arbeiten. Damit ist gemeint, dass jeder Krankheit ein bestimmtes Symptomenmuster fix zugeordnet wird. Ein sehenswertes Beispiel dafür ist CONSULTANT

(http://www.vet.cornell.edu/consultant/consult.asp). In der Tabelle wird beispielsweise der Krankheit K1 das Symptomenmuster S1, S3 und S5 zugeordnet.

Die Diagnose muss jedoch natürlich von den Symptomen ausgehen. In dem Beispiel aus Abbildung 2.4-2 sähe die Diagnostik wie folgt aus: Da S1 vorliegt, scheidet K2 aus. Aufgrund der Anwesenheit von S2 kommt nur noch K3 in Frage.

In manchen gut strukturierbaren Gebieten der Medizin lassen sich Diagnostik-Programme nach dieser Weise erarbeiten.

Es ist jedoch leicht einzusehen, dass es sich bei diesem Verfahren um eine Vereinfachung handelt.

Jede Vereinfachung hat ihren Preis! Der Preis, der für die Vereinfachung der Verarbeitung bezahlt werden muss, ist der Informationsverlust, der durch die Reduktion von ursprünglich quantitativen Merkmalen zu qualitativen entsteht.

2.4.2 Symptome als konstante, quantitative Merkmale

Jeder bearbeitete Fall einer Krankheit wird durch einen Messwert repräsentiert.

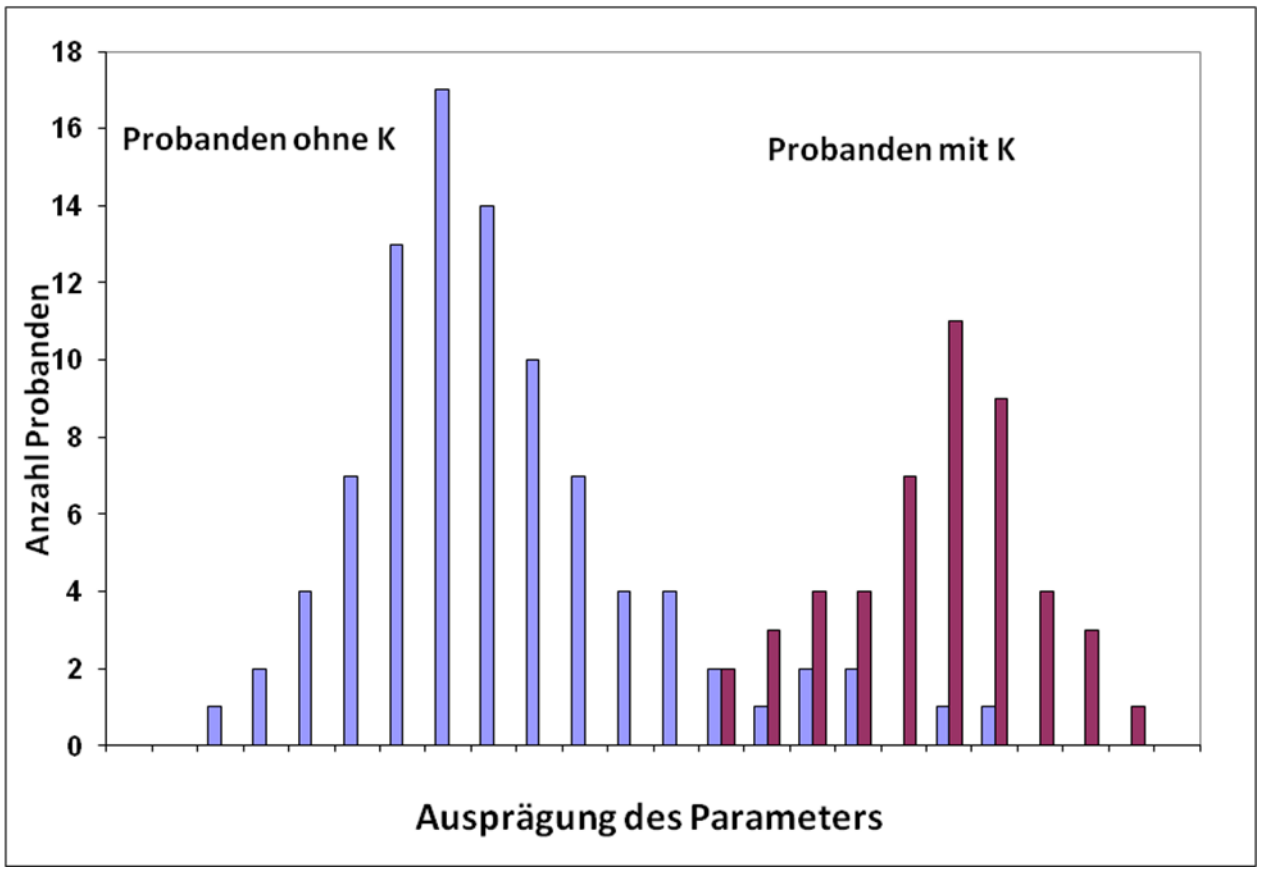


Abb. 2.4-3: Ausprägung eines Parameters bei Individuen mit und ohne Krankheit K

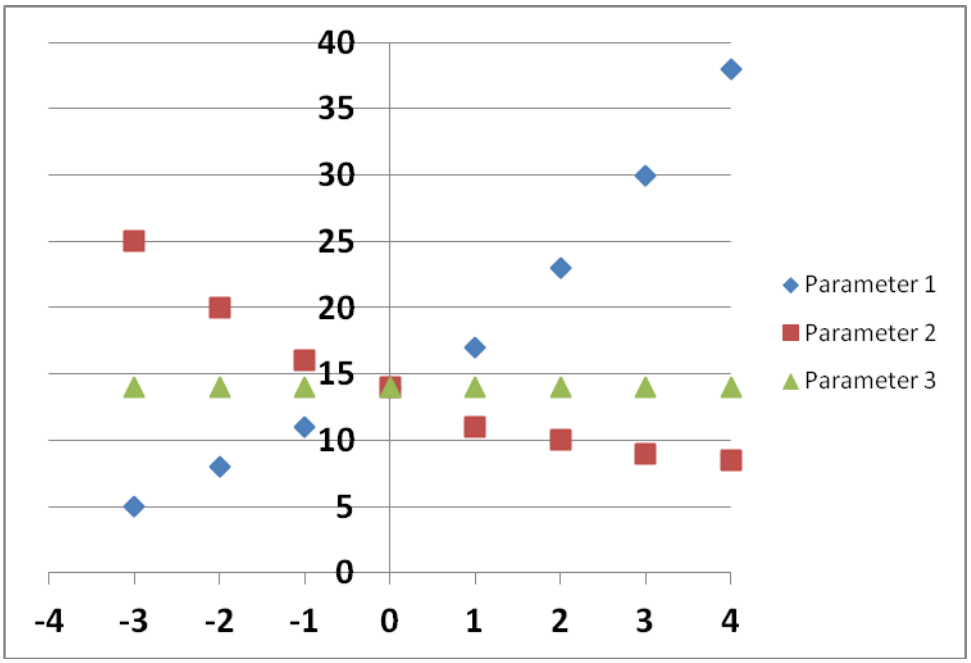
Die Ordinate gibt dabei die Anzahl der Fälle an, bei denen der jeweilige Messwert (Abszisse) festgestellt wurde (s. Abb. 2.4-3).

Es ist offensichtlich, dass sich eine solche Grafik leicht erstellen lässt, dass es aber in der Realität kaum möglich sein wird, die dazu nötigen Daten zu bekommen.

Die formalisierte Verarbeitung ist auch hier relativ einfach. Sie besteht in der Aufstellung einer so genannten **ROC-Kurve** (receiver-operating characteristics). Der Gewinn, der durch den größeren Aufwand erkauft wird, besteht darin, dass stärkere Abweichungen vom Referenzintervall zu größerer Sicherheit in der Diagnose führen. Die ROC-Kurve erlaubt es auch, zwischen verschiedenen Prioritäten zu wählen (vgl. Abschnitt über ROC-Kurven 2.7).

2.4.3 Symptome als dynamisch quantitative Merkmale

Ein zu einem bestimmten Zeitpunkt gemessener Wert kann auf verschiedene Weise zustande gekommen sein (s. Abb. 2.4-4).



Zeit

Abb. 2.4-4: Verschiedene Möglichkeiten der Ausprägung von Parametern im Verlauf der Zeit vor und nach der Bestimmung zum Zeitpunkt 0

Die formalisierte Verarbeitung eines Verlaufs ist schwierig, es sei denn, Verläufe werden wie konstante Merkmale (also zum Beispiel stetiger Anstieg, anhaltende, leicht fluktuierende Erhöhung etc.) behandelt.

2.5 „Normalwerte“ und Referenzintervalle

Der Begriff „Normalwerte“ wurde und wird in der Medizin häufig verwendet. Normalität ist ein vieldeutiger Begriff und wird oft als Synonym für Gesundheit angesehen. Für „Gesundheit“ gibt es verschiedene Definitionen (unter anderem eine viel zitierte der WHO aus dem Jahr 1948: „Der Zustand vollkommenen physischen, psychischen und sozialen Wohlbefindens, nicht lediglich die Abwesenheit von Krankheit“), die aber so allgemein gehalten sind, dass sie zur Bearbeitung konkreter Fragestellungen unbrauchbar sind. Die Feststellung von Gesundheit ist in der Tiermedizin am ehesten bei Ankaufsuntersuchungen relevant.

Normalwerte wurden zum Teil in recht willkürlicher Weise erstellt (in einem konkreten Beispiel durch Untersuchung von Blutproben von fünf freiwilligen Soldaten). Der Begriff „Normalwerte“ wird mehr und mehr durch den Begriff „Referenzintervall“ ersetzt. Für die Erstellung von Referenzintervallen gibt es sehr ausführliche Anweisungen (z.B. Gräsbeck, R., u. T. Alström Hrsg.: Reference Values in Laboratory Medicine. The Current State of the Art). (1981) So müssen Referenzintervalle unter anderem der Fragestellung und den Umständen angepasst werden. Wenn zum Beispiel die Erhöhung der Atemfrequenz von Jungmastbullen in Laufställen als früher Indikator einer Atemwegserkrankung verwendet werden soll, wird man als Probanden zur Erstellung des Referenzintervalls nicht Kühe auf der Weide nehmen. Wenn man es genau nimmt, gibt es ein Problem, das entweder eine Zirkeldefinition oder einen potentiell unendlichen Rekurs erfordert, und sich in etwa so formulieren lässt: Zur Erstellung des Referenzintervalls der Atemfrequenz von Jungmastbullen ohne Erkrankung des Atemapparates sind solche Rinder auszuwählen, deren Atemfrequenz im Referenzintervall liegt, oder deren Atemapparat gesund ist, wobei die Gesundheit des Atmungsapparates anhand anderer Parameter festgelegt werden muss, für welche die analoge Problematik besteht. Für die klinische Praxis spielt diese Problematik aber zum Glück keine sehr große Rolle.

Werte innerhalb eines Referenzintervalls werden als unauffällig, solche außerhalb als auffällig und erklärungsbedürftig, mithin als Indikation für weitergehende Diagnostik angesehen. Umgekehrt werden Werte im Referenzintervall als Hinweis dafür angesehen, dass das zugehörige Organsystem oder die zugehörige Körperfunktion vermutlich nicht der Sitz einer bestehenden Erkrankung ist.

Ob Abweichungen nach oben und nach unten als auffällig anzusehen sind, zum Beispiel bei der Glukosekonzentration im Blut, oder nur nach einer Seite, wie zum Beispiel bei der Aktivität des Enzyms Kreatinkinase im Serum oder Plasma, hängt von der klinischen Relevanz ab.

Auf das Kapitel „Referenzintervalle“ im Skriptum „Klinische Pathophysiologie und Labordiagnostik“ wird hingewiesen.

Für die Differentialdiagnostik, also die Frage, ob Krankheit K oder eine andere Krankheit vorliegt, spielen Referenzintervalle meist nur in der Weise eine Rolle, dass manche Krankheiten weitestgehend ausgeschlossen werden können, wenn bei ihnen ein Parameter stets verändert ist, also im Extremfall eine diagnostische Sensitivität von 1,0 (und damit auch einen prädiktiven Wert des negativen Testausfalls von 1,0) hat, bei einem Patienten jedoch im Referenzintervall ist. Ein Beispiel wäre ein negativer Ausfall des Glutartests bei der Frage, ob

eine Kuh Endokarditis hat. Hohe diagnostische Spezifität von Abweichungen von einem Referenzintervall kommt nicht so häufig vor, am ehesten bei Parametern, die in der Definition der Erkrankung verwendet werden oder „nahe dran“ sind. Beispiele wären die Erniedrigung der Kalziumkonzentration im Blut (Hypokalzämie, Gebärparese) und die Erhöhung des so genannten TTP-Effektes (Thiaminmangel).

2.6 Trennwerte

Testergebnisse werden häufig als "positiv" oder "negativ" bzw. als "normal" oder "pathologisch/auffällig" dargestellt. Eine solche Beschränkung auf zwei Ergebnisklassen nennt man dichotom. Sie kommt den Bedürfnissen des Klinikers entgegen, denn eine Intervention kann meist nur ganz oder nicht durchgeführt werden, nicht aber zu 87 %.

Dichotome Testergebnisse sind meist pseudoqualitativ, d.h. eigentlich quantitativ, und es gibt einen Trennwert, auf Englisch cutoff point oder cutoff value, anhand dessen ein Ergebnis beurteilt oder eingeordnet wird, wie in der folgenden Abbildung 2.6-1 dargestellt.

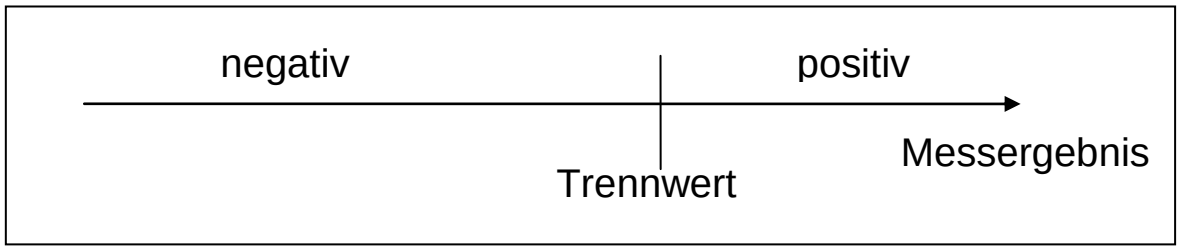


Abb. 2.6-1: Einteilung von Messergebnissen in „positiv“ und „negativ“

Zur Festlegung von Trennwerten siehe Kapitel „Referenzintervalle“ im Skriptum „Klinische Pathophysiologie und Labordiagnostik“.

Die Reduktion eines quantitativen Ergebnisses auf ein qualitatives ist mit einem Gewinn an "Handhabbarkeit" verbunden. Mit anderen Worten: mit einer Vereinfachung. Die Testergebnisse können leichter verarbeitet werden. Aber jeder Gewinn, jeder Vorteil, hat seinen Preis. Der Preis besteht hier in einem Verlust an Information, nämlich der Information darüber, wie weit der individuelle Messwert vom Trennwert entfernt liegt.

Die Lage des Trennwerts ist von großer Bedeutung. Wenn sich die Wertebereiche von Individuen mit einer Krankheit K und derjenigen ohne diese Krankheit mehr oder weniger stark überschneiden, wie in der Abbildung 2.6-2 gezeigt, bestimmt die Lage des Trennwertes die Sensitivität und Spezifität des Tests.

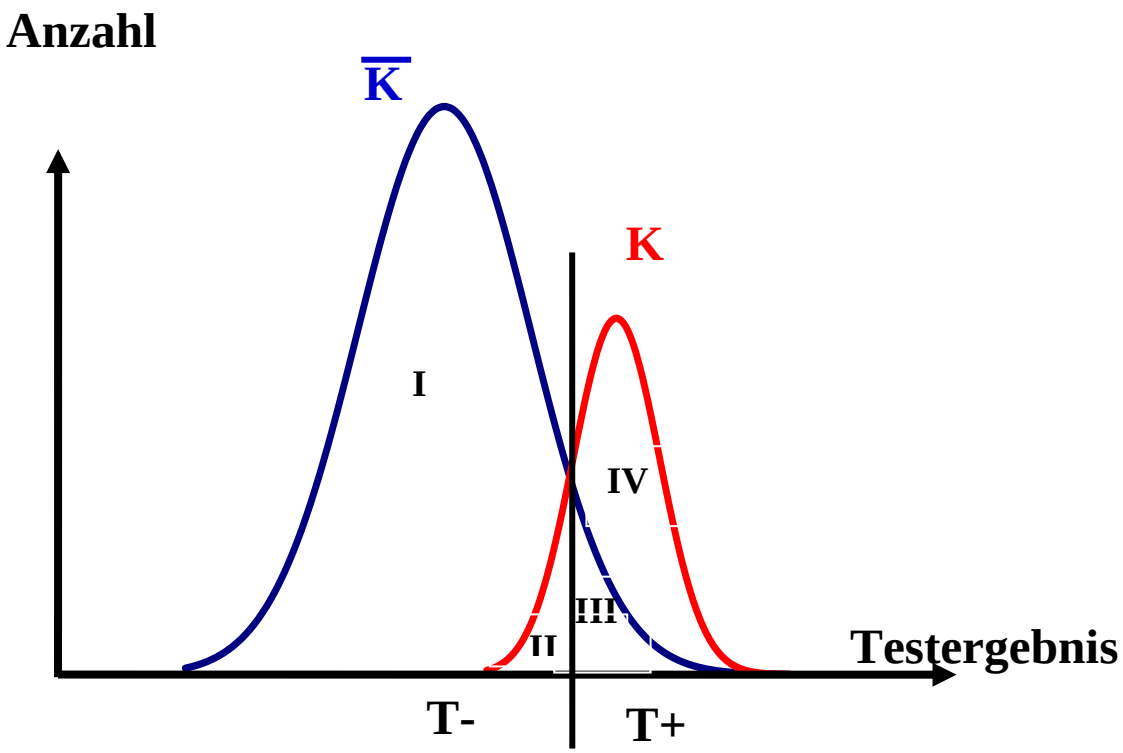


Abb. 2.6-2 Trennwert für kontinuierliche Parameter

Frage 2.6-1: Welchen Feldern der Vierfeldertafel entsprechen die mit I bis IV bezeichneten Flächen unter den Kurven in Abb. 2.6-2?

Frage 2.6-2: Welche Auswirkungen hat eine Verschiebung des Trennwertes nach rechts oder nach links auf Sensitivität und Spezifität des Tests?

In den Abbildungen 2.6-3 und 2.6-4 wird das Problem anhand zweier Fischpopulationen erläutert. Die (unerwünschten) „Weißlinge“ sind im Durchschnitt größer als die (erwünschten) Rötlinge, aber die beiden Populationen überlappen sich in der Größe.

Sollen Weißlinge herausgefangen werden, gibt es zwei Möglichkeiten: entweder nimmt man im Bestreben, möglichst alle Weißlinge zu fangen (hohe Sensitivität), ein relativ feinmaschiges Netz, mit der Folge, dass auch Rötlinge gefangen werden (geringe Spezifität) oder man nimmt im Bestreben, möglichst nur Weißlinge zu fangen (hohe Spezifität) ein Netz mit gröberen Maschen, mit der Folge, dass auch mehr oder weniger viele Weißlinge „durch die Maschen“ gehen (geringe Sensitivität).

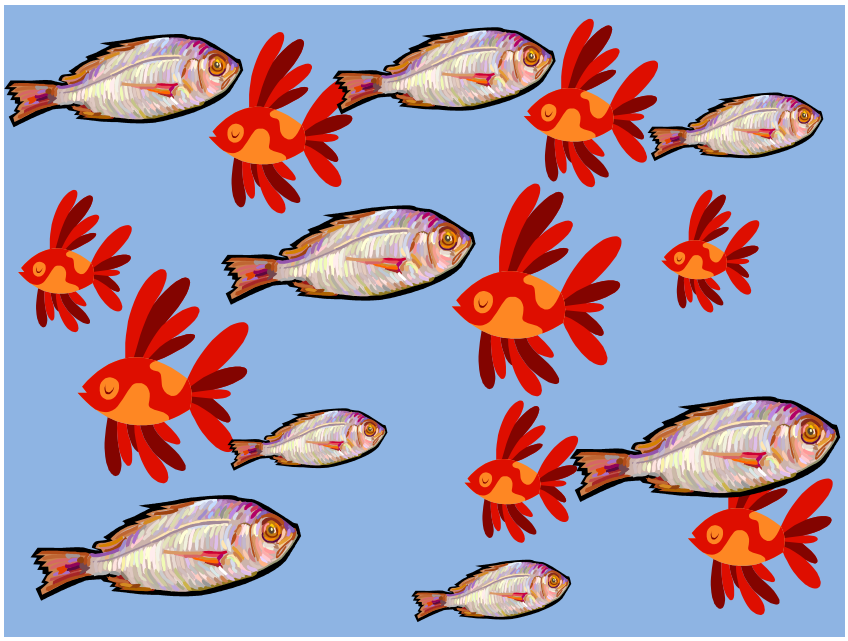


Abb. 2.6-3: Zwei Fischpopulationen, die sich in der Größe überlappen

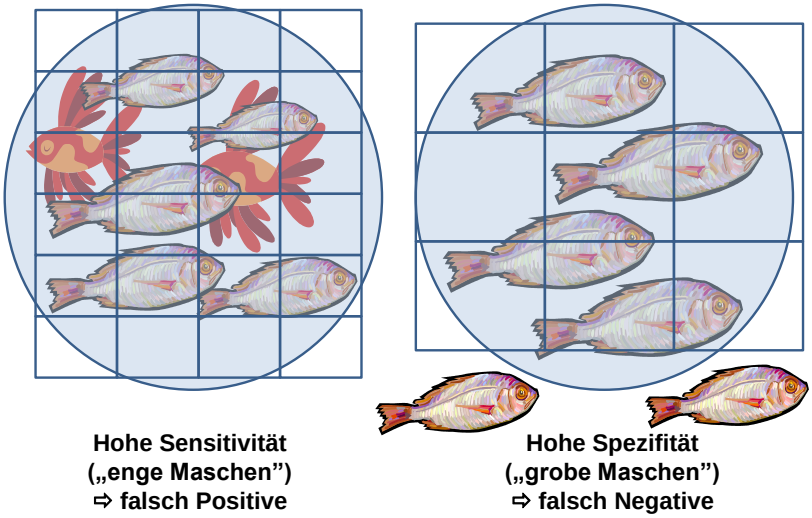


Abb. 2.6-4: Tausch zwischen Sensitivität und Spezifität in überlappenden Populationen

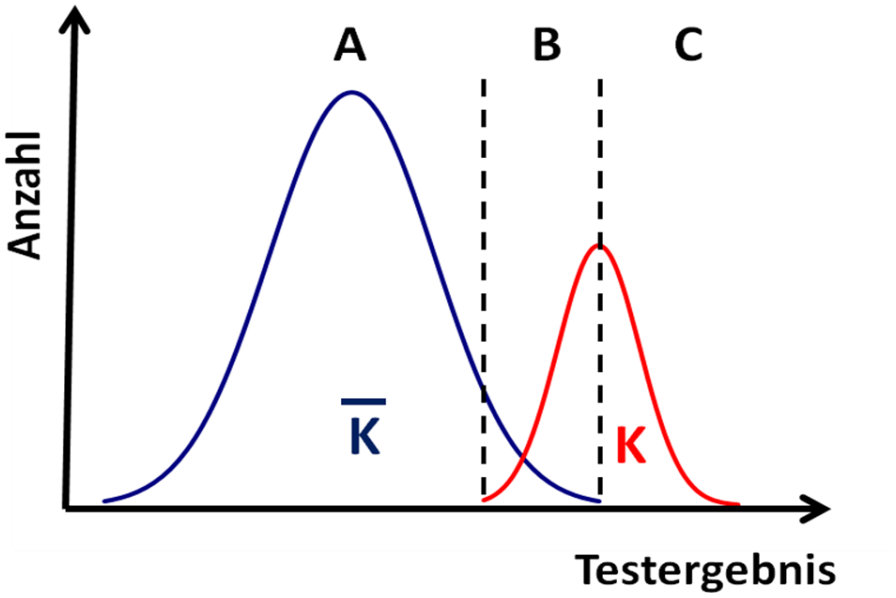


Abb. 2.6-5: Aufteilung des Wertebereichs in drei Teile

In Abb. 2.6-5 sind die Verteilungskurven der Testergebnisse bei Individuen ohne und mit K dargestellt. Dabei gibt es offensichtlich drei Teile (A, B und C), die sich hinsichtlich der Interpretation deutlich unterscheiden. Ergebnisse im Teil A schließen die Krankheit K aus, während Krankheit K bei Ergebnissen im Teil C als nachgewiesen betrachtet werden kann. Ergebnisse im Teil (B) können nicht eindeutig interpretiert werden. Hier besteht die Indikation zu weiteren Untersuchungen. Die Prävalenz von K spielt dabei keine wichtige Rolle, denn Lage und Spannweite der Werte bleiben gleich.

In diesem Zusammenhang lässt sich ein interessantes Phänomen besprechen. Zu Beginn der Bekämpfung der Rindertuberkulose nach dem zweiten Weltkrieg in Westdeutschland war der Anteil falsch positiver Ergebnisse im intrakutanen Tuberkulintest bezogen auf die gesamte Rinderpopulation (!) relativ konstant und niedrig. Deshalb war er zunächst zu vernachlässigen, weil die Prävalenz von Rindertuberkulose viel höher war. Durch die Tilgungsmaßnahmen nahm aber die Prävalenz von Rindertuberkulose rasch drastisch ab, so dass schließlich die Individuen mit falsch positivem Ergebnis (das meist auf die Infektion mit so genannten atypischen Mykobakterien oder anderen antigenetisch verwandten Keimen zurückzuführen war) einen immer größeren Anteil der Reagenten ausmachten. Ihre Tilgung verursachte weiterhin jedes Jahr Kosten, ohne dass damit aus seuchenhygienischer Sicht Vorteile verbunden waren. Die Kosten für jeden entdeckten echten Fall stiegen an. Dieses Problem, das bei der Bekämpfung jeder Infektionskrankheit besteht, ist in Abbildung 2.6-6 dargestellt.

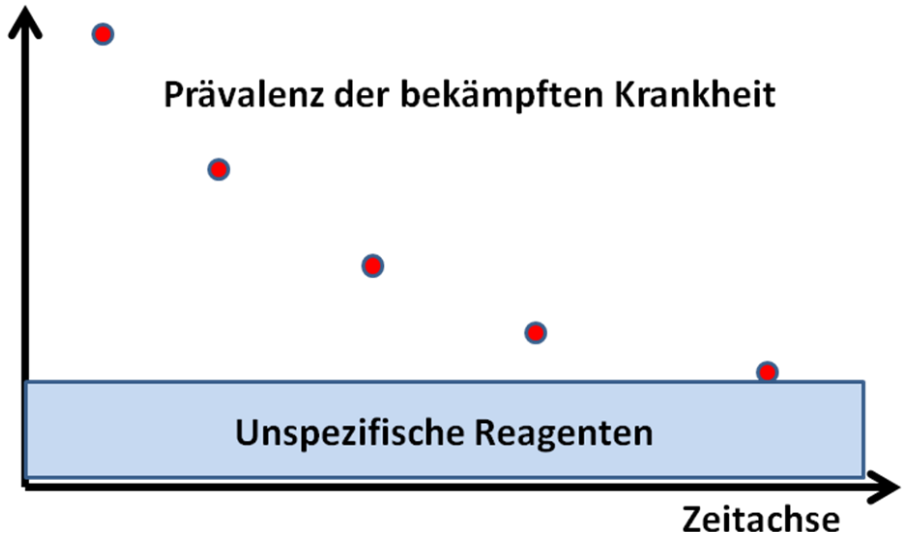


Abb. 2.6-6: Relation der Prävalenz von echten Fällen und unspezifischen Reagenten im Verlauf der Bekämpfung einer Infektionskrankheit.

Frage 2.6-3: Welche Möglichkeit gibt es zur Abhilfe, wenn davon ausgegangen wird, dass Rinder mit Tuberkulose im Allgemeinen stärker reagieren als Rinder, welche mit den oben genannten anderen Keimen infiziert sind?

2.7 Receiver Operating Characteristics (ROC)-Kurven

Der etwas merkwürdig anmutende Begriff wurde ursprünglich im II. Weltkrieg zur Beschreibung der Fähigkeit und Einstellung (= „Operation Characteristics“) von Funkern oder Radaroperateuren (= „Empfängern“) zur Unterscheidung von Signal und Rauschen geprägt. Die formalisierte Verarbeitung in Form von ROC-Kurven lässt sich aber auf allgemeine Probleme der Unterscheidung zweier Teil-Populationen anwenden.

So kann die Berücksichtigung des **Grades der Abweichung** eines konkreten Ergebnisses von diesem Trennwert in diagnostischer Sicht verarbeitet werden.

Ein Beispiel aus der Humanmedizin: Die Senkung der ST-Strecke im EKG nach Belastung kann zur Diagnostik von koronarer Herzkrankheit oder Koronarinsuffizienz herangezogen werden. Bei einem Kollektiv von Patienten wurde nach dem Belastungs-EKG eine Katheterisierung der Herzkranzgefäße vorgenommen, und die Ergebnisse wurden als „Goldstandard“ zur Beurteilung der EKG-Befunde verwendet. Daraus ergab sich folgende Tabelle:

Tab. 2.7-1: Sensitivität und Spezifität in Abhängigkeit von verschiedenen „cutoff values“ für die ST-Senkung im Belastungs-EKG (aus: Forrester et al. Circulation 65 (1981) 915-921)

Ergebnis	Sensitivität	Spezifität
≥ 0,5 mm	0,86	0,62
≥ 1,0 mm	0,65	0,85
≥ 1,5 mm	0,42	0,96
≥ 2,0 mm	0,33	0,98
≥ 2,5 mm	0,20	0,995

Für verschiedene Trennwerte sind die sich ergebende diagnostische Sensitivität und Spezifität aufgeführt. Wie zu erwarten, nimmt die diagnostische Sensitivität umso mehr ab, je strenger das Kriterium ist, je höher sozusagen die Messlatte für die Diagnose „koronare Herzkrankheit“ gelegt wird. Gleichzeitig steigt die diagnostische Spezifität an. Ein immer geringerer Anteil der Individuen ohne Koronarerkrankung zeigt eine ST-Absenkung, welche über den angenommenen Trennwert hinausgeht.
Anhand der aufgeführten Werte lässt sich die in Abbildung 2.7-1 dargestellte Kurve konstruieren. Sie wird „performance characteristics“-Kurve oder „reveiver-operating-characteristics“-Kurve oder kurz ROC-Kurve genannt. In ihr kommt u.a. die schon angesprochene Abhängigkeit der Sensitivität und Spezifität eines Tests von der Lage des Trennwertes zum Ausdruck

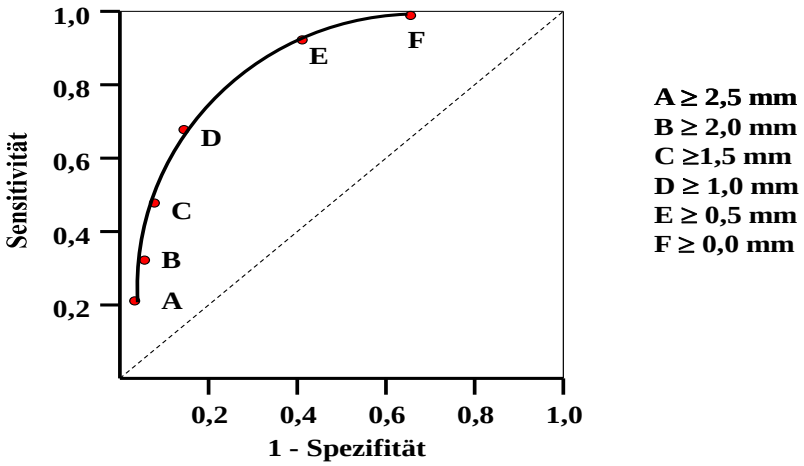


Abb. 2.7-1: ROC-Kurve für die Validität der ST-Senkung im Belastungs-EKG zur Diagnostik von koronarer Herzerkrankung

Es ist zu beachten, dass die Höhe der auf der Abszisse aufgetragenen Spezifität von rechts nach links zunimmt!

Es gibt auch andere Darstellungsweisen für ROC-Kurven. Sie unterscheiden sich aber nicht prinzipiell von der hier gewählten.

Frage 2.7-1: Welche Bedeutung hat die Winkelhalbierende in dieser Darstellung?

Wenn sich die Testwerte zweier Teilpopulationen teilweise überlappen, sind Sensitivität und Spezifität gegenläufig, können also nicht gleichzeitig erhöht werden. Das kommt in dieser Darstellung zum Ausdruck. Beim Vergleich zweier Tests in einer bestimmten Situation kann dagegen ohne weiteres ein Test sowohl sensitiver als auch spezifischer sein als der andere (vgl. Abb. 2.7-2).

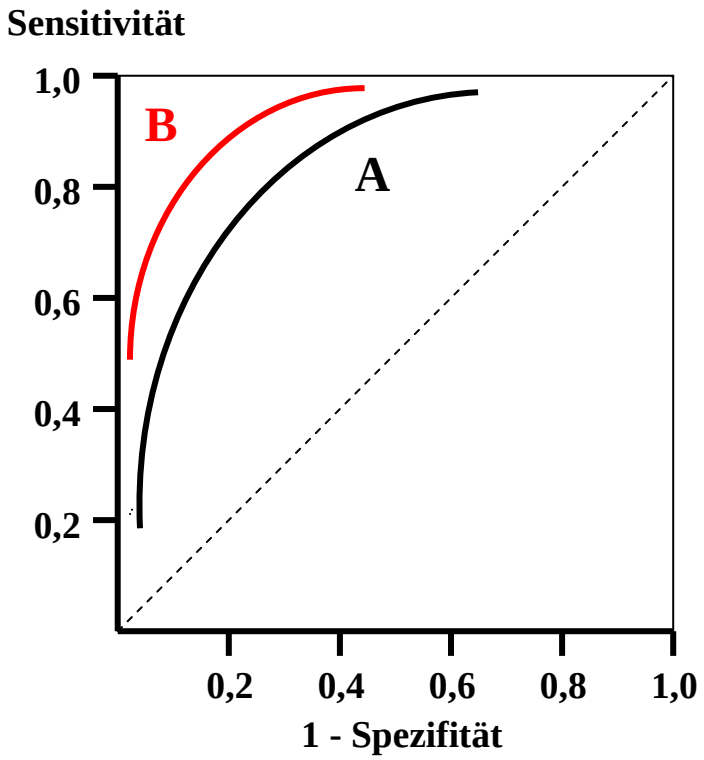


Abb. 2.7-2: Idealisierte ROC-Kurven zweier Tests A und B

Frage 2.7-2: Welcher Test ist in dieser Situation besser?

Die Lage der Kurve, die einen Test repräsentiert, hinsichtlich der Ordinaten, drückt also aus, wie gut der Test zwischen den beiden Zielgruppen zu unterscheiden vermag.

Frage 2.7-3: Wie sähe die ROC-Kurve des idealen Tests in dieser Darstellungsweise aus?

In einer Untersuchung über die Qualität der Beurteilung von Röntgenbildern nach den Kriterien "pathologisch" oder "nicht pathologisch" durch verschiedene erfahrene Röntgenologen (Goldstandard: mehrheitliche Meinung einer Gruppe von Experten) zeigte sich, dass ihre Resultate auf verschiedenen Stellen einer ROC-Kurve lagen.

Frage 2.7-4: Was bedeutet das?

Es gibt verschiedene quantitative Verfahren zum Vergleich verschiedener Tests. Eines besteht darin, die Flächen unter den ROC-Kuren zu vergleichen. Je größer diese Fläche ist, umso besser ist der Test. Eine andere Möglichkeit zum Vergleich ist die maximale Länge des Lotes

von der Kurve auf die Winkelhalbierende. Der Schnittpunkt dieses Lotes mit einer ROC-Kurve markiert die Kombination von Sensitivität und Spezifität mit dem maximalen Anteil richtiger Zuordnungen (= maximale Effektivität). Das ist in Abbildung 2.7-3 für die Daten aus Tabelle 2.7-1 dargestellt. Der optimale Cutoff liegt danach zwischen 0,5 und 1,0 mm ST-Absenkung mit einer dSE von etwa 0,8 und einer dSp von etwa 0,75. Die Summen aus den exakten Werten für dSe und dSp sind sowohl bei 0,5 mm als auch bei 1,0 mm ST-Absenkung kleiner.
Ein weiterer Parameter zum Vergleich von Tests ist der Youden-Index γ. Er ist definiert als
γ = Sensitivität – (1 – Spezifität).

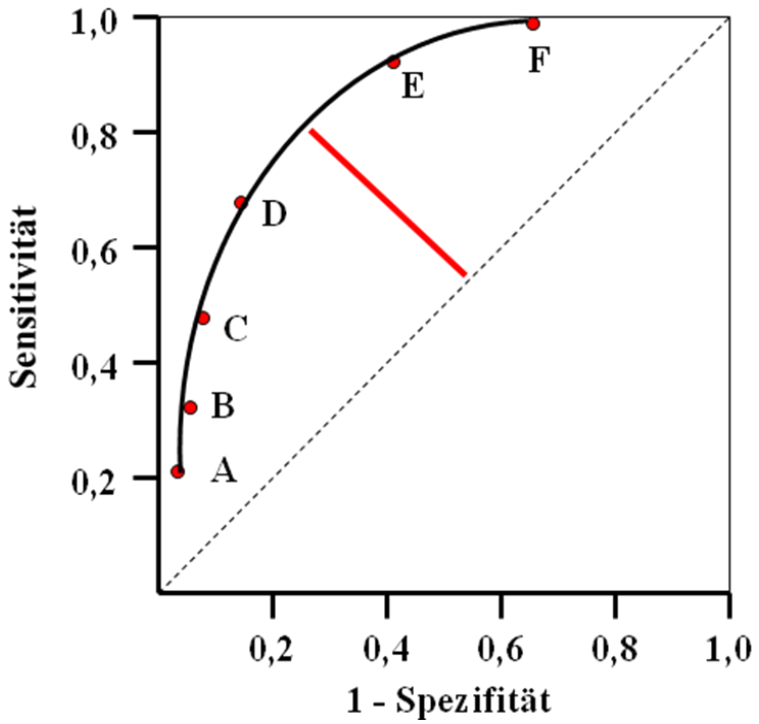


Abb. 2.7-3: Bestimmung des Cutoffs mit dem höchsten Anteil korrekter Zuordnungen

Eine weitere Möglichkeit könnte darin bestehen, die Anteile richtiger und falscher Zuordnungen an den Punkten maximaler Richtigkeit mithilfe des Chi-Quadrat-Tests zu vergleichen. Bei der Frage, ob eine Test signifikant besser als ein anderer ist, muss beachtet werden, dass statistische Signifikanz nicht nur von der Größe des Unterschieds, sondern auch von der Größe der Stichprobe abhängt.

Nicht berücksichtigt werden dabei bisher die relativen Kosten falsch positiver und falsch negativer Zuordnungen. Sind diese bekannt, kann der Trennwert (und damit Sensitivität und Spezifität) entsprechend eingestellt werden.

Der optimale Trennwert liegt da, wo die Steigung der Kurve folgendem Wert entspricht:

Steigung der ROC-Kurve am optimalen Trennwert
= Nettokosten FP * p(1-K)/(Nettokosten FN * p(K)) **(10)**

Netto-Kosten FP = Kosten FP – Kosten RN
Netto-Kosten FN = Kosten FN – Kosten RP

Aus dieser Formel geht u.a. hervor: bei geringer Apriori-Wahrscheinlichkeit [p(K)] liegt der optimale cutoff point an einer steilen Stelle, d.h.: die Spezifität muss erhöht werden. Das gleiche gilt, wenn die Nettokosten FP deutlich größer sind als die Nettokosten FN.
Soviel als Einführung in die ROC-Kurven-Analyse.

2.8 Goldstandard

Zur Bestimmung der diagnostischen Sensitivität und Spezifität ist es unerlässlich, den wahren Status der Probanden unabhängig vom Ausfall des Tests zu ermitteln. Das ist das Problem des so genannten "gold standard", wie es in der amerikanischen Literatur genannt wird, also die Entscheidung darüber, anhand welcher Kriterien ein Erkrankungsfall als „ein Fall von Krankheit K“ eingeordnet wird. Falldefinitionen sind in der epidemiologischen und klinischen Literatur von großer Bedeutung, denn von ihnen hängt u.a. ab, ob Fälle aus verschiedenen Publikationen mit einander verglichen werden können. Es gibt verschiedene Möglichkeiten der Statusermittlung, wenn die Krankheit im Prinzip oder tatsächlich pathologisch-anatomisch definiert ist, was häufig der Fall ist.

2.8.1. Retrospektive Auswertung "gesicherter Fälle"

Probleme: "Gesicherte Fälle" bedeuten bei einer pathologisch definierten Krankheit in der Regel tote Patienten. Daher besteht ein "Bias" (also ein systematischer Fehler mit gerichteter Beeinflussung des Ergebnisses) zugunsten schwerwiegender Fälle.
Es können auf diese Weise nur Angaben zur diagnostischen Sensitivität von zuvor erhobenen Befunden, nicht aber zu deren diagnostischer Spezifität, gewonnen werden.

2.8.2. Verfolgen der (Verdachts-)Fälle einer klinisch definierten Situation („clinical setting)“ bis zum bitteren Ende

Eine solche klinisch definierte Situation könnte zum Beispiel lauten „Kuh mit intermittierendem Durchfall). Im Prinzip ist das ein Bias-freier Ansatz; nicht jedoch, wenn die fragliche Krankheit "narbenfrei" verheilen kann. Sonst erfolgt wiederum Selektion schwerer Verlaufsformen. Theoretisch wären so Angaben sowohl zur diagnostischen Sensitivität als auch zu (klinisch relevanten) diagnostischen Spezifität erhobener Befunde möglich. Die so ermittelten Schätzwerte für die diagnostische Sensitivität und Spezifität gelten für das Patientengut der Institution, an der sie ermittelt wurden. Dieses Verfahren würde auch das Problem umgehen, das sich aus der Verwendung von gesunden Individuen zur Bestimmung der Spezifität ergibt.

2.8.3. Vergleich mit den Ergebnissen einer Referenzmethode

In der Nierenfunktionsdiagnostik ist die Inulin-Clearance nach dem Standard-Verfahren ein solcher Referenztest für die glomeruläre Filtrationsrate (GFR).
Problem: Der zu prüfende Test kann nie besser sein als die Referenzmethode.
Eine ggf. bestehende höhere Sensitivität würde als höhere Quote an falsch positiven Ergebnissen, also als geringere Spezifität gewertet, eine höhere Spezifität als eine höhere Quote an falsch negativen Ergebnissen, also als geringere Sensitivität. Eine "Wahrheit hinter der Wahrheit" kann nur erahnt, nicht aber bewiesen werden.

Im Rahmen der Besprechung des Goldstandards lässt sich ein anderes Problem „unterbringen“. Warum ist es oft so schwierig, die richtige Diagnose zu stellen?

Hier ist ein kurzer Ausflug in die Geschichte recht aufschlussreich. Bis vor etwa 200 Jahren war Schwindsucht eine Diagnose, die von den Ärzten oft gestellt wurde. Die Diagnose war relativ einfach, denn zum einen handelte es sich um eine recht häufige Erscheinung und zum anderen bedurfte es zur Stellung dieser Diagnose nur der Erhebung des Vorberichtes und der

klinischen Untersuchung. Die Definitionskriterien dieser „Krankheit“ lagen auf der Ebene der klinischen Symptomatik und waren daher dem Diagnostiker direkt zugänglich.

Für die frühen Pathologen waren die grobsinnlichen Befunde, die sie an den Leichen von Schwindsüchtigen erhoben, jedoch keineswegs einheitlich. Mit dem Einzug der Mikroskopie in die Pathologie wurde entdeckt, dass bei vielen dieser Menschen ganz ähnliche histologische Veränderungen vorlagen. Das traf aber auch auf Individuen zu, die nicht an Schwindsucht gestorben waren. Und damit begann sozusagen das Leid der Kliniker. Denn nun rutschten die Definitionskriterien auf die Ebene der Pathologie, mit dem Ergebnis, dass ganz unterschiedliche klinische Bilder der neuen Krankheit Tuberkulose zugeordnet werden mussten, was naturgemäß nicht mit absoluter Sicherheit möglich war. Nach Entdeckung und Differenzierung der Mykobakterien als Erreger der Tuberkulose rutschten die Definitionskriterien erneut einen Stock tiefer, dieses Mal in die Ebene der Ätiologie. Damit verbunden war, dass die durch *Mycobacterium tuberculosis* verursachte Krankheit des Menschen auch solche Formen beinhaltete, die nicht mit der Bildung von spezifischem Granulationsgewebe (Tuberkel) verbunden war, so zum Beispiel eine akute Sepsis. Es ist einleuchtend, dass damit eine weitere Erschwerung der klinischen Diagnostik verbunden war.

In Grafik 2.8-1soll diese Problematik illustriert werden. Der Einfachheit halber sind nur die drei Ebenen Klinik, Pathologie und Ätiologie dargestellt. Dabei repräsentiert eine kleine rote Ellipse die Definitionskriterien einer Krankheit, die grauen Ellipsen in den Ebenen darunter und/oder darüber die damit verbundenen Phänomene auf den jeweiligen Ebenen.

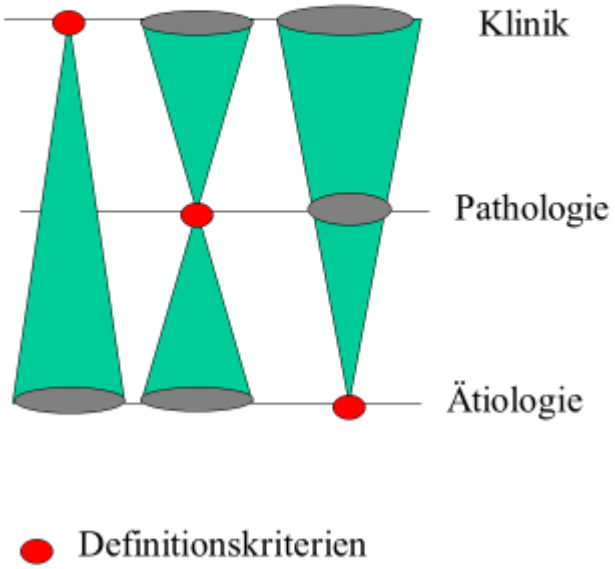


Abb. 2.8-1: Zusammenhang zwischen Ätiologie, pathologischer natomie und klinischer Symptomatik (Erläuterung siehe Text)

Bestimmte, auf einer Ebene definierte Phänomene sind jeweils mit einem Spektrum von Phänomenen auf den anderen beiden Ebenen verbunden. Deshalb sind meist nur Wahrscheinlichkeitsschlüsse von klinischen Symptomen auf das Vorliegen von pathologisch oder ätiologisch definierten Krankheiten möglich.

Dieses Schema verdeutlicht auch, dass es kein nosologisches System (Krankheitsklassifikation) geben kann, das die beiden Grundforderungen an Klassifikationen erfüllt, nämlich Vollständigkeit und Eindeutigkeit, solange die Definitionskriterien verschiedener Krankheiten auf verschiedenen Ebenen liegen.

"Definitionen" von Krankheiten in Lehrbüchern sind meist keine regelrechten Definitionen im Sinne von Krankheit A liegt genau dann vor, wenn die Bedingungen a bis n erfüllt sind“, sondern eher Kurzbeschreibungen.

2.9 Das Bayes-Theorem

Thomas Bayes (1702 – 1761) war ein englischer Geistlicher und Mathematiker. Aus seinem Nachlass wurde 1763 eine Arbeit veröffentlicht: "Essay towards solving a problem in the doctrine of chances" (Transactions of the Royal Society 53, 370 ff). Diese Arbeit befasst sich mit der konkreten Verarbeitung probabilistischer Information. Genauer gesagt, sie befasst sich damit, wie sehr wir vorhandenes Wissen (Apriori-Wahrscheinlichkeit) nach Erlangung zusätzlicher Information zu einer Aposteriori-Wahrscheinlichkeit modifizieren müssen. Das davon abgeleitete so genannte BAYES-Theorem ist sehr berühmt geworden und ist Grundlage vieler computer-unterstützter Diagnostik-Systeme.

Es gibt eine Reihe von Darstellungsformen dieses Theorems. Eine der einfachsten ist folgende:

$$p(K/S) = p(K) * p(S/K)/p(S) \qquad (11)$$

p(K) ist die Apriori-Wahrscheinlichkeit für Krankheit K. Ohne sonstige Information entspricht sie der Prävalenz.
p(S/K) ist die bedingte Wahrscheinlichkeit für das Symptom S unter der Voraussetzung, dass Krankheit K vorliegt, also die diagnostische Sensitivität von S für K.
p(S) ist die Häufigkeit des Symptoms S in der fraglichen Population. Je größer dieser Wert, also je häufiger das Symptom, umso geringer ist offensichtlich die Spezifität des Symptoms S für die Krankheit K.
p(K/S) ist die bedingte Wahrscheinlichkeit für die Krankheit K unter der Voraussetzung, dass Symptom S vorliegt.
Eine weitere Darstellung ist
Prävalenz*Sensitivität/(Prävalenz*Sensitivität + (1-Prävalenz)*(1-Spezifität))

Das BAYES-Theorem kann auch für beliebig viele Symptome und Krankheiten formuliert werden und ist dann naturgemäß viel komplizierter.

Frage 2.9-1: Der letzte Satz ist offensichtlich nur eine andere Formulierung für die Definition des prädiktiven Wertes. Wie kann man zeigen, dass die beiden Formeln äquivalent sind?

2.10 Übereinstimmung von Testergebnissen (Kappa-Statistik)

In bestimmten Situationen kann es sinnvoll sein zu fragen wie gut Tests oder Untersucher übereinstimmen. Stellen wir uns vor, zwei Untersucher würden unabhängig voneinander 100 Kälber mittels der Schwingauskultation untersuchen, also prüfen, inwieweit Plätschergeräusche zu provozieren sind, wenn die Bauchwand in Schwingungen versetzt wird.

Die Ergebnisse dieser Untersuchung sind in Tabelle 2.10-1 eingetragen:

Tab. 2.10-1: Beurteilung der Schwingauskultation (SA) bei 100 Kälbern durch zwei Untersucher (A und B)

		Untersucher A		
		SA positiv	SA negativ	
Untersucher B	SA positiv			53
	SA negativ			47
		54	46	100

Die Übereinstimmung zwischen den beiden Untersuchern erscheint fast vollkommen, wenn wir die globalen Ergebnisse betrachten. Ist sie es aber wirklich? Aus der Tabelle geht nicht hervor, wie die Übereinstimmung bei den einzelnen Patienten war. Betrachten wir Tabelle 2.10-2.

Tab. 2.10-2: Beurteilung der Schwingauskultation (SA) bei 100 Kälbern durch zwei Untersucher (A und B)

		Untersucher A		
		SA positiv	SA negativ	
Untersucher B	SA positiv	40	13	53
	SA negativ	14	33	47
		54	46	100

Nun sieht die Sache etwas anders aus. Übereinstimmung besteht in 40 + 33 = 73 von 100 Fällen, also in 73 %. Wir müssen uns fragen, ob denn diese Übereinstimmung nicht weitgehend zufällig ist. Aber wie viel an Übereinstimmung wäre auf Grund des Zufalls zu erwarten? Wie weiter oben schon ausgeführt, ist die Wahrscheinlichkeit des gemeinsamen Auftretens voneinander unabhängiger Ereignisse gleich dem Produkt der Einzelwahrscheinlichkeiten. Bezogen auf unser Beispiel bedeutet dies Folgendes: Die Wahrscheinlichkeit, dass Untersucher A bei einem der Kälber einen positiven Ausfall der Schwingauskultation registriert, beträgt 54/100 = 0,54. Bei Untersucher B beträgt die entsprechende Wahrscheinlichkeit 0,53. Die Wahrscheinlichkeit für die zufällige Übereinstimmung hinsichtlich des positiven Ausfalls der Schwingauskultation beträgt 0,54 * 0,53 = 0,29, und diejenige für den negativen Ausfall 0,46 * 0,47 = 0,22. Insgesamt wäre also auf Grund des Zufalls eine Übereinstimmung in 0,29 + 0,22 = 0,51 (also 51 %) der Fälle zu erwarten. Die tatsächliche Übereinstimmung in Höhe von 0,73 ist höher.

Als Maß für die „überzufällige“ Übereinstimmung wurde die Größe κ (COHEN'S kappa) eingeführt. Sie gibt an, welcher Anteil der möglichen „überzufälligen“ Übereinstimmung tatsächlich erreicht wurde. In unserem Beispiel beträgt die zufällig zu erwartende Übereinstimmung 0,51, was für „überzufällige“ Übereinstimmung 0,49 „übrig“ lässt. Davon haben die Untersucher 0,22 (0,73– 0,51) erreicht, also 45 % oder 0,45. κ ist hier 0,45.
Die Höhe der möglichen zufälligen (und damit auch der „überzufälligen“) Übereinstimmung ist nicht konstant, sondern hängt von der Höhe der Einzelwahrscheinlichkeiten ab. Diese nicht sofort einleuchtende Tatsache soll anhand Tabelle 2.10-3 verdeutlicht werden.

Tab. 2.10-3: Beurteilung der Schwingauskultation (SA) bei 100 Kälbern durch zwei Untersucher (C und D)

		Untersucher C		
		SA positiv	SA negativ	
Untersucher D	SA positiv	66	12	78
	SA negativ	15	7	22
		81	19	100

Auch hier stimmen die Ergebnisse der beiden Untersucher in 73 % der Fälle überein. Die allein auf Grund des Zufalls zu erwartende Übereinstimmung errechnet sich, wie oben dargelegt, mit

$$0{,}81*0{,}78 + (1-0{,}81)*(1-0{,}78) = 0{,}63 + 0{,}19*0{,}22 = 0{,}63 + 0{,}04 = 0{,}67.$$

Die Übereinstimmung zwischen Untersucher C und D liegt also nur 6 %-Punkte[1] (0,73 – 0,67 = 0,06) über der zufällig zu erwartenden. Das entspricht einem Anteil von 18 % (0,18) des für die „überzufällige“ Übereinstimmung verbleibenden Restes von 0,33. κ ist hier bei gleicher absoluter Übereinstimmung wie im ersten Beispiel (73 % oder 0,73) lediglich 0,18. Das bedeutet, dass die mögliche Größe von κ von der Prävalenz abhängt.

Frage 2.10-1: Wann ist die zufällig zu erwartende Übereinstimmung am kleinsten?

Die Frage, ob sich die Übereinstimmung zwischen den Ergebnissen zweier Untersucher (oder Tests) in statistisch signifikanter Weise von der zufällig zu erwartenden unterscheiden, kann auf verschiedene Weise bearbeitet werden. Zum einen kann ein χ^2-Test (Chi-Quadrat-Test) durchgeführt werden. Verwenden wir dazu das Beispiel aus Tabelle 2.10-2.

Tab. 2.10-4: Erwartete und beobachtete Übereinstimmung zweier Untersucher

	Zufall	Untersucher A und B
Übereinstimmung	51	73
Keine Übereinstimmung	49	27

Der Unterschied in der Besetzung der Felder ist statistisch signifikant.

[1] Bei der Verwendung von Prozenten sollte man sehr genau darauf achten, dass die Bezugsgröße eindeutig ist.

Eine andere Möglichkeit zur Beurteilung besteht in der Berechnung des 95 %-Vertrauensbereiches für κ: $\kappa - 1{,}96\ s(\kappa)$ bis $\kappa + 1{,}96\ s(\kappa)$. Hierbei ist $s(\kappa)$ die (geschätzte) Standardabweichung von κ. Wenn dieser Vertrauensbereich Null nicht einschließt, heißt das, dass die Übereinstimmung signifikant größer als die zufällig zu erwartende ist.

Eine dritte Methode ist die Bestimmung eines so genannten u-Werts, der sich aus dem Quotienten aus kappa und seiner geschätzten Standardabweichung berechnet.

Die Kappa-Statistik kann auch für Befunde mit mehr als zwei Ausprägungen verwendet werden. Zwei Untersucher, A und B, beurteilen unabhängig voneinander die Farbe der Maulschleimhäute bei 100 Kuh-Patienten einer Klinik und ordnen sie in die Kategorien „gerötet“, „unauffällig“ oder „blass“ ein. Tab. 2.10-5 zeigt das tatsächliche Ergebnis, während in Tab. 2.10-6 das nach dem Zufall zu erwartende Ergebnis wiedergegeben ist.

Tab. 2.10-5: Übereinstimmung zweier Untersucher bei Befunden mit drei Ausprägungen

		Untersucher A			
		gerötet	unauffällig	blass	
Untersucher B	gerötet	12	12	0	24
	unauffällig	6	40	5	51
	blass	1	10	14	25
		19	62	19	100

Tab. 2.10-6: Aufgrund des Zufalls zu erwartende Übereinstimmung zweier Untersucher bei Befunden mit drei Ausprägungen

		Untersucher A			
		gerötet	unauffällig	blass	
Untersucher B	gerötet	4,56	14,88	4,56	24
	unauffällig	9,69	31,62	9,69	51
	blass	4,75	15,50	4,75	25
		19	62	19	100

Die tatsächliche Übereinstimmung beträgt 12 + 40 + 14 = 66 von 100, also 66 %. Wie die Zellen in Tabelle 2.10-6 ausgerechnet werden, soll anhand des Beispiels der Zelle demonstriert werden, in welcher die Anzahl der Patienten steht, bei denen Untersucher A den Befund „blass“, Untersucher B dagegen den Befund „“unauffällig“ erhoben hat. In dieser Zelle steht in Tabelle 2.10-5 die Zahl 5. Die Zahl 9,69, die an dieser Stelle in Tabelle 2.10-6 steht, kommt dadurch zustande, dass die beiden Randsummen (19 bzw. 51) miteinander multipliziert und das Ergebnis durch die Gesamtsumme (100) dividiert wird.
Wie aus Tabelle 2.10-6 hervorgeht, beträgt die rein zufällig zu erwartende Übereinstimmung 4,56 + 31,62 + 4,75 = 40,93 von 100, also 40,93 %. Die weitere Berechnung von Kappa erfolgt wie oben.

Frage 2.10-2: Wie lautet das Ergebnis?

Wie die Betrachtung der Tabelle 2.10-5 zeigt, gab es zwischen den beiden Untersuchern zwar eine Reihe von Abweichungen in der Befundung, aber bis auf einen Fall (in dem die Farbe der

Schleimhaut von Untersucher A als gerötet, von Untersucher B jedoch als blass eingestuft wurde) betrafen sie eine Stufe. Es gibt die Möglichkeit, ein gewichtetes Kappa zu berechnen, wobei die Stärke der Abweichungen berücksichtigt wird. Hierzu wird auf Lehrbücher der Epidemiologie und Biomathematik verwiesen.

2.11 Diagnostik-Methoden

Da Epidemiologen oft mit Daten arbeiten, die von Klinikern erhoben wurden, sollen im Folgenden einige Diagnostik-Methoden kurz besprochen werden. Es besteht eine bemerkenswerte Diskrepanz zwischen dem, was Kliniker an Diagnostikmethode lehren und dem, was sie tatsächlich bei ihrer täglichen Arbeit tun. Die Lehre lässt sich mit Sprüchen wie „Die gründliche klinische Untersuchung ist die Grundlage tierärztlichen Handelns“ (Richard Götze) zusammenfassen. Es wird gefordert, zunächst unvoreingenommen eine (ungezielte) gründliche und „vollständige“ klinische Untersuchung durchzuführen und dann die erhobenen Befunde kritisch auszuwerten. Diese Methode könnte „**Exhaustionsmethode**“ genannt werden, weil „erschöpfend“ untersucht wird. Soweit die Theorie. In der Praxis gehen Kliniker jedoch völlig anders vor. Auf der Basis des Vorberichts und einiger Nachfragen stellen sie innerhalb weniger Sekunden eine Liste der häufigsten Krankheiten auf, die für die geschilderten Symptome verantwortlich sein könnten, und dann gehen sie in ihrer Untersuchung gezielt vor, um die Liste „abzuarbeiten“. „Die klinische Untersuchung“ (so lautet ein Teil des Titels eines Lehrbuches der buiatrischen Propädeutik) gibt es also in der Praxis nicht, sondern klinische Untersuchungsmethoden, die je nach Lage des Falles indiziert sind und daher alle beherrscht werden sollten. Man stelle sich vor, was ein Landwirt sagen würde, wenn ein Tierarzt, den er wegen der Euterentzündung einer Kuh gerufen hat, die Kuh einer kompletten Untersuchung einschließlich Klauenkontrolle, rektaler Untersuchung sowie Pansensaft- und Liquorentnahme unterziehen würde. Als besondere Aspekte in der Nutztiermedizin könnten zur Rechtfertigung einer „vollständigen“ Untersuchung angeführt werden, dass sich dabei Gesichtspunkte ergeben können, welche von erheblicher prognostischer Bedeutung sind, im Einzelfall also dazu führen können, dass aus wirtschaftlichen oder seuchenhygienischen Gründen von einer Behandlung der eigentlichen Krankheit abgesehen wird (Beispiel: eine vom Besitzer als hochträchtig angesehene Kuh, die wegen Euterwarzen vorgestellt wird, stellt sich bei der rektalen Untersuchung als nicht trächtig heraus). Zudem können sich bei der Untersuchung Hinweise auf Umstände ergeben, die für den Gesamtbestand von Bedeutung sind (Beispiel: Reheringe an den Klauen oder Befall mit Ektoparasiten). Außerdem besagt ein weiterer Spruch zu Recht, dass mehr Fehldiagnosen auf Mängeln in der Untersuchung als auf Mängeln im Wissen beruhen. All diese Überlegungen ändern aber nichts an der Tatsache, dass Kliniker in der oben beschriebenen Weise vorgehen, also sehr früh eine begrenzte Anzahl von diagnostischen „Hypothesen“ bilden, welche sie im Verlauf der Untersuchung „testen“. Diese Art der Diagnostik wird „**hypotheto-deduktiv**“ genannt. Allerdings gehen Kliniker dabei weniger in der Art von Naturwissenschaftlern vor, welche Experimente ersinnen, deren Ergebnisse geeignet sein können, eine Hypothese zu „falsifizieren“, sondern sie suchen meist nach Hinweisen, welche eine Hypothese unterstützen. Möglicherweise besteht ein grundsätzlicher Unterschied zwischen Naturwissenschaft und praktischer (Tier-)Medizin darin, dass sich die Richtigkeit diagnostischer „Hypothesen“ im Gegensatz zu derjenigen naturwissenschaftlicher Art tatsächlich belegen lässt.

Bei manchen Krankheiten sind, wenn man sie einige Male gesehen hat, so genannte Blickdiagnosen möglich,. Ein Beispiel ist in Abb. 2.10-1 dargestellt.

Abb. 2.11-1 Beispiel für eine durch „Blickdiagnose“ zu erkennende Krankheit (Trichophytie**)**

Blickdiagnosen sind eine spezielle Form der **Mustererkennung** (engl.: pattern recognition). Auch sie beruhen auf der Erkennung von Ähnlichkeiten. Die Fähigkeit, Ähnlichkeiten zu erkennen, ist angeboren. Sie ist identisch mit der Fähigkeit, einer Klasse zuzuordnen, also sozusagen Dinge in eine „Schublade“ zu tun. Nicht angeboren ist dagegen das Klassifikationssystem, also Zahl und Bezeichnung der „Schubladen“. Ähnlichkeit zu erkennen ist eine alltägliche Erfahrung, sie klar zu definieren, ist dagegen sehr schwer und vielleicht nicht sinnvoll. Es ist uns mitunter nicht möglich zu sagen, worauf der Eindruck der Ähnlichkeit beruht. Der Mechanismus der Mustererkennung läuft also auf einer Ebene „unterhalb“ des Bewusstseins ab. Daher kann diese Art der Diagnostik auch nur sehr schlecht gelehrt werden. Es gibt keinen Ersatz für die persönliche und unmittelbare Erfahrung des Gesamtbildes, der „Gestalt“ von Patienten.

Wenn Ärzte oder Tierärzte diagnostische Maßnahmen delegieren, entwerfen sie oft schriftliche Anweisungen in Form von Flussdiagrammen, die ein klares Vorgehen erlauben. Manche diagnostischen Aufgaben lassen sich auf diese Weise gut lösen. Ein Beispiel ist in Abb. 2.11-2 wiedergegeben. Diese Art der Diagnostik kann „**logische Verzweigung**“ genannt werden.

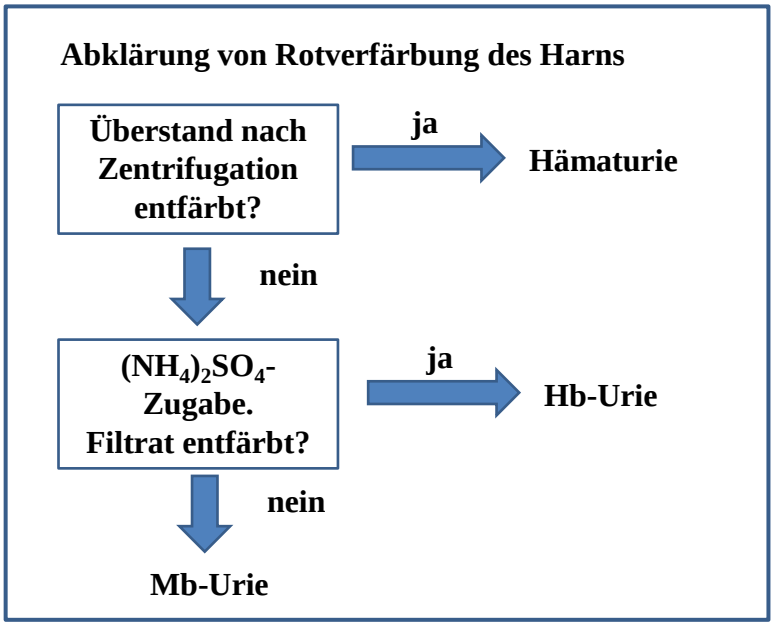


Abb. 2.11-2: Beispiel für Diagnostik nach Art der logischen Verzweigung

Problemorientierte Diagnostik: Aus Vorbericht, klinischer und Labor-Untersuchung wird eine „Problem-Liste“ erstellt. Für jedes Problem werden möglichen Krankheiten („rule-outs“) aufgelistet. Die Krankheit, welche die meisten Probleme (Symptome) „erklärt“, oder aber diejenige, die am dringendsten der Intervention bedarf, wird bevorzugt weiterverfolgt. Das erste Prinzip entspricht „Ockhams Rasiermesser“, wonach zur Erklärung eines Phänomens nicht unnötig viele Hypothesen aufgestellt werden sollten.

2.12 Diagnostik in Populationen

Bisher war fast immer das Individuum, der einzelne Patient, Gegenstand unseres diagnostischen Interesses. Informationen aus Gruppen (Prävalenz, Sensitivität, Spezifität) wurden nur verwendet, um die Validität von Tests im Bezug auf den einzelnen Fall zu bewerten. Es können jedoch auch Gruppen (Herden, Populationen unterschiedlicher Größe) Gegenstand des diagnostischen Interesses sein. In Tabelle 2.12-1 sind einige der dabei möglichen Fragestellungen im Vergleich zu denen der „Individualdiagnostik“ aufgeführt.

Tab. 2.12-1: Gegenüberstellung von Fragestellungen der Individualdiagnostik und der Diagnostik in Populationen

Einheit des Interesses	
Individuum	**Population**
Sensitivität, Spezifität und prädiktive Werte von qualitativen Attributen *Beispiel: Spezifität des Befundes „Verzögerung des Lidreflexes“ für D-Laktat-Azidose*	Anwesenheit oder Abwesenheit oder Prävalenz von qualitativen Attributen *Beispiel: Anteil von Ausscheidern von M. avium subsp. paratuberculosis in einer Milchkuhherde*
Höhe und Dynamik quantitativer Attribute *Beispiel: Konzentration von L-Laktat im Blut bei einer Kuh mit Labmagentorsion*	Mittlere Größe und Verteilung von quantitativen Attributen *Beispiel: Mittlere Hämoglobinkonzentration in einer Gruppe von Milchmastkälbern* Anteil der Population mit Werten innerhalb/außerhalb eines Bereichs *Beispiel: Anteil von Kühen mit mehr als 200.000 Zellen pro ml Milch.* Abklärung von Krankheitsausbrüchen *Beispiel: Auftreten tödlich verlaufender hämorrhagischer Diathese bei jungen Kälbern (BNP) in verschiedenen Ländern*
	Optimale Auswahlstrategie bei regionalen Untersuchungen auf Seuchenfreiheit

2.12.1 Anwesenheit oder Abwesenheit von qualitativen Attributen in einer Population

Unter Kälbern war die Prävalenz von persistierender Infektion (PI) mit dem Virus der bovinen Virusdiarrhoe (BVDV) vor Beginn der staatlichen Bekämpfung etwa 3 %. (Die genaue Zahl spielt hier keine entscheidende Rolle.) Diese Tiere scheiden das Virus ständig massiv aus und sind daher eine Ansteckungsquelle für andere. Ein Bullenmäster, der regelmäßig Gruppen von 30 Kälbern zukauft, fragt seinen Tierarzt, wie groß das Risiko (= die Wahrscheinlichkeit) ist, in einer solchen Gruppe mindestens ein PI-Kalb zu haben. (Es wird angenommen, dass eine zugekaufte Gruppe eine repräsentative Stichprobe der Kälberpopulation darstellt, also jedes männliche Kalb der Gesamtpopulation die gleiche Chance hat, in der Gruppe zu landen.)

Die Antwort kann offensichtlich nicht 3 % und auch nicht 30*3 % heißen. Sie ist dagegen relativ leicht zu finden, wenn man sozusagen das Pferd von hinten aufzäumt und sich überlegt, wie groß die Wahrscheinlichkeit ist, dass ein Kalb kein PI-Tier ist. Diese Wahrscheinlichkeit muss 100 – 3 = 97 % (also 0,97) betragen. Die Wahrscheinlichkeit, dass die ersten beiden Kälber keine PI-Tiere sind, ist daher 0,97*0,97 = $0,97^2$ = 0,94, denn die Wahrscheinlichkeit des gleichzeitigen Auftretens zweier voneinander unabhängiger Ereignisse ist gleich dem Produkt der beiden Einzelwahrscheinlichkeiten. Die Wahrscheinlichkeit, dass alle 30 Kälber keine PI-Tiere sind, beträgt demnach $0,97^{30}$ = 0,40. Hinter den übrigen 60 % verbergen sich alle Kombinationen von einem bis zu 30 PI-Tieren. Die Formel zur Berechnung der ursprünglichen Fragestellung lautet also:

Wahrscheinlichkeit für Anwesenheit eines qualitativen Attributes in einer Stichprobe = 1 – (1 – Prävalenz des Attributes)$^{\text{Stichprobengröße}}$ (12)

Zur Illustration sind in Tabelle 2.12.-2 die Wahrscheinlichkeiten für einige Prävalenzen und Stichprobengrößen aufgeführt.

Tab. 2.12-2: Wahrscheinlichkeit [1 – (1 – Präv.)^n] des Nachweises eines qualitativen Attributes in einer Stichprobe in Abhängigkeit von der Prävalenz des Attributes in der Population und der Stichprobengröße

	Anzahl untersuchter Tiere (n)							
Prävalenz	**5**	**10**	**20**	**25**	**30**	**50**	**75**	**100**
0,01	0,049	0,096	0,182	0,222	0,260	0,395	0,529	0,634
0,02	0,096	0,183	0,332	0,397	0,455	0,636	0,780	0,867
0,03	0,141	0,263	0,456	0,533	0,599	0,782	0,898	0,952
0,04	0,185	0,335	0,558	0,640	0,706	0,870	0,953	0,983
0,05	0,226	0,401	0,642	0,723	0,785	0,923	0,979	0,994
0,06	0,266	0,461	0,710	0,787	0,844	0,955	0,990	0,998
0,07	0,304	0,516	0,766	0,837	0,887	0,973	0,996	0,999
0,08	0,341	0,566	0,811	0,876	0,918	0,985	0,998	1,000
0,09	0,376	0,611	0,848	0,905	0,941	0,991	0,999	
0,1	0,410	0,651	0,878	0,928	0,958	0,995	1,000	
0,12	0,472	0,721	0,922	0,959	0,978	0,998		
0,14	0,530	0,779	0,951	0,977	0,989	0,999		
0,16	0,582	0,825	0,969	0,987	0,995	1,000		
0,18	0,629	0,863	0,981	0,993	0,997			
0,2	0,672	0,893	0,988	0,996	0,999			
0,25	0,763	0,944	0,997	0,999	1,000			
0,3	0,832	0,972	0,999	1,000				
0,35	0,884	0,987	1,000					
0,4	0,922	0,994						
0,45	0,950	0,997						
0,5	0,969	0,999						

Die völlige Abwesenheit eines qualitativen Attributes in einer Population, also z.B. die Freiheit von einem Erreger, kann mit absoluter Sicherheit nur durch mehrmalige

Untersuchung aller Individuen mit einem sehr sensitiven Verfahren nachgewiesen werden. In der Praxis begnügt man sich daher mit Untersuchungen, mit denen eine bestimmte Prävalenz mit einem vorgegebenen Grad an Sicherheit nachgewiesen werden kann.

Die oben genannte Formel kann auch umgeformt werden, wenn die Größe einer Stichprobe bestimmt werden soll. Deutschland gilt derzeit als frei von Rinderleukose. Dieser Status wird anerkannt, wenn mindestens 99,8 % der Bestände leukosefrei sind. Zur Aufrechterhaltung dieses Status ist jährlich eine repräsentative Stichprobe von Betrieben per Tankmilchserologie zu untersuchen, deren Größe es erlaubt, 0,2 % infizierte Betriebe mit einer Sicherheit von 95 % zu erfassen.
Ausgehend von Formel (12) ergibt sich
$1 - 0{,}998^n = 0{,}95$
$0{,}998^n = 0{,}05$
$n*\ln 0{,}998 = \ln 0{,}05$
$n = \ln 0{,}05 / \ln 0{,}998 = 1496$

Die Vorstellung, eine in einer Population ermittelte Prävalenz eines qualitativen Merkmals, z.B. Antikörper gegen *Mycobacterium avium* subsp. *paratuberculosis*, könne in gleicher Weise auf jedes Individuum der Population projiziert werden, ist oft nicht korrekt. Die spezielle Epidemiologie einer Infektionskrankheit darf nicht vernachlässigt werden. So ist in einer infizierten Herde die Wahrscheinlichkeit für die Anwesenheit von Antikörpern bei den ältesten Rindern wesentlich höher als bei Jungrindern.

2.12.2 Abschätzung der Prävalenz eines qualitativen Attributes mit vorgegebener Präzision und Sicherheit.

Bei manchen Infektionskrankheiten hängt die Art der zu wählenden Bekämpfungsstrategie (beispielsweise Impfung oder Merzung der ganzen Herde) vom Anteil betroffener Tiere in der Herde (= Prävalenz) ab. Daher ist es wichtig zu wissen, wie hoch die Prävalenz ist. Bei kleineren Herden können alle Tiere untersucht werden, während das in großen Herden meist nicht praktikabel ist. Die Größe der zu untersuchenden Stichprobe hängt von verschiedenen Faktoren ab. Es ist leicht einzusehen, dass umso mehr Tiere untersucht werden müssen, je genauer wir das Ergebnis haben wollen, je geringer also die Abweichung der geschätzten Prävalenz von der wahren sein soll. Außerdem hängt die Stichprobengröße von der Sicherheit ab, die wir haben wollen, dass die wahre Prävalenz tatsächlich in dem durch die gewählte Präzision begrenzten „Fenster“ liegt. Und letztlich hängt die Größe der Stichprobe auch noch von der Prävalenz selbst ab.

Die Prävalenz soll mit einer Präzision von L = ± 5 % und einer Sicherheit von 95 % geschätzt werden. Die vermutete Prävalenz wird mit P^ („P Dach“; das Zirkumflex sollte über dem P stehen) bezeichnet. Gibt es keinerlei Anhaltspunkte über die Prävalenz, ist sie mit 50 % (0,5) anzusetzen.

Die allgemeine Formel lautet:

$$\mathbf{n = 3{,}84*P\hat{}*(1-P\hat{})/L^2} \qquad \mathbf{(13)}$$

Die Zahl 3,84 ist das Quadrat der Zahl 1,96, welche den Vertrauensbereich von 95 % repräsentiert. (Unter einer Normalverteilungskurve enthält der Bereich zwischen Mittelwert – 1,96 Standardabweichungen und Mittelwert + 1,96 Standardabweichungen 95 % der Gesamtfläche unter der Kurve.)
Setzt man die oben angegebenen Zahlen ein, ergibt sich als nötiger Umfang der Stichprobe

$n = 3,84*0,5*(1-0,5)/0,05^2 = 384$

In Abbildung 2.12.-1 ist die Abhängigkeit des Stichprobenumfangs von der geschätzten Prävalenz bei einer Präzision von ± 5 % für die Vertrauensbereiche 95 % und 99 % dargestellt. Wie sich zeigt, ist der Umfang der nötigen Stichprobe bei einer geschätzten Prävalenz von 0,5 am größten.

Ein konkretes Beispiel: Während der Blauzungenkrankheit-Epidemie wurden Verordnungen erlassen, in denen festgelegt wurde, dass zur Erkennung der Blauzungenkrankheit bei empfänglichen Wildtieren Untersuchungen durchzuführen sind, um mit einer Wahrscheinlichkeit von 95 % und einer angenommenen Prävalenz von 0,5 % befallene Tiere zu erkennen.
Prävalenz = 0,005

$(1-Pr)^n = 0,05$

$n*\log(1-Pr) = \log(0,05)$

$n= (\log(0,05)/(\log(0,995))$

$n = 598$

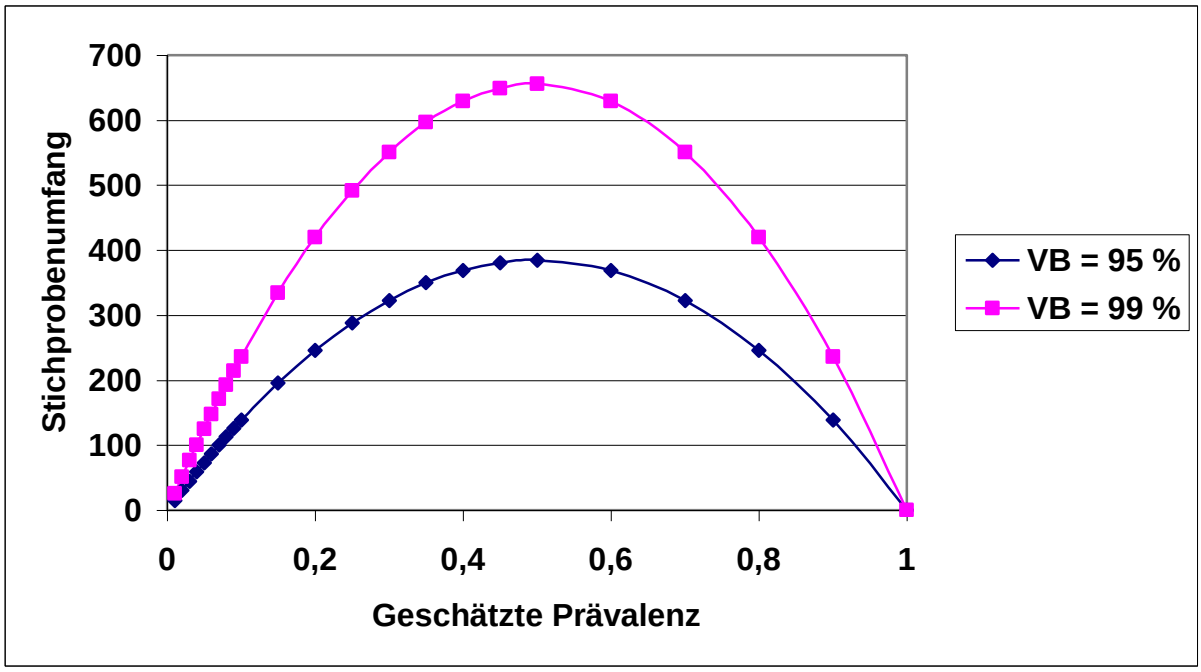


Abb. 2.12-1: Umfang von Stichproben zur Ermittlung der Prävalenz eines qualitativen Attributes bei Präzision von ± 5 % und Vertrauensbereich 95 % oder 99 % in Abhängigkeit der geschätzten Prävalenz

Zur Berechnung des Umfangs von Stichproben in kleineren Herden gibt es eine Formel:

1/korrigiertes n = 1/errechnetes n + 1/Herdengröße

Unter Verwendung der oben genannten Bedingungen ergibt sich hierbei für eine Herde von 30 Tieren folgender Stichprobenumfang:
1/korr. n = 1/384 + 1/30 = 28

Das ergibt offensichtlich keinen großen Gewinn gegenüber einer Untersuchung der Gesamtherde. Aber wir haben auch bei maximaler Unsicherheit (P^ = 0,5) große Ansprüche an die Präzision (L = ± 5 %) gestellt. Könnten wir zuversichtlich sein, dass die wahre Prävalenz um 20 % liegt, und wären wir mit einer Präzision von ± 10 % zufrieden, ergäbe sich bei VB = 95 % ein Stichprobenumfang von 61 Tieren in großen Herden und ein solcher von 20 in einer Herde von 30 Tieren.

Die Tatsache, dass die Unsicherheit bei einer geschätzten Prävalenz von 50 % am größten ist, mag nicht unmittelbar einleuchten. Daher soll sie durch eine Grafik erläutert werden. In Abbildung 2.12-2 sind drei „Populationen“ mit unterschiedlicher Prävalenz des qualitativen Attributs „roter voller Kreis“ schematisch dargestellt. Im linken Feld beträgt die Prävalenz 5 %, im mittleren 50 % und im rechten 95 %. Wenn wir jeweils eine Reihe von 10 Individuen als eine Zufallsstichprobe betrachten, dann ist die Streuung der Ergebnisse ganz offensichtlich im mittleren Feld am größten, denn sie reicht von 2/10 bis 8/10, während sie im linken Feld von 0/10 bis 2/10, im rechten sogar nur von 9/10 bis 10/10 reicht. Die Streuung hat aber einen großen Einfluss auf die Größe der zu erfassenden Stichprobe.

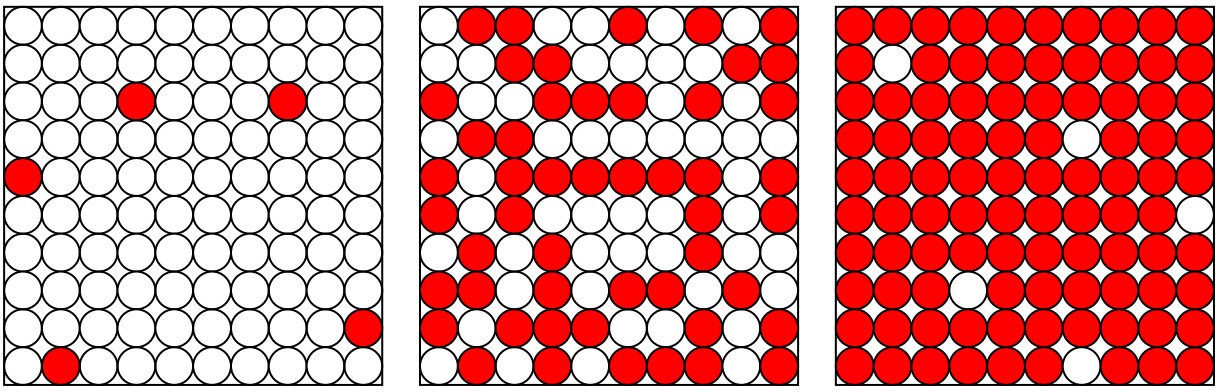


Abb. 2.12-2: Schematische Darstellung unterschiedlicher Häufigkeiten (5 %, 50 % und 95 %) eines qualitativen Attributs in drei verschiedenen Populationen. (Rote gefüllte Kreise symbolisieren betroffene Individuen.)

Im Folgenden soll anhand einer größeren „Population“ demonstriert werden, wie groß der Vertrauensbereich für die Schätzung der Prävalenz eines qualitativen Attributes in Abhängigkeit der Größe der Stichprobe ist.
In Abbildung 2.12-3 sind 400 Kreise, welche eine Population repräsentieren sollen. Anhand von Zufallszahlen wurden 40 davon ausgewählt. Die Prävalenz des Attributs beträgt also 0,1 oder 10 %.

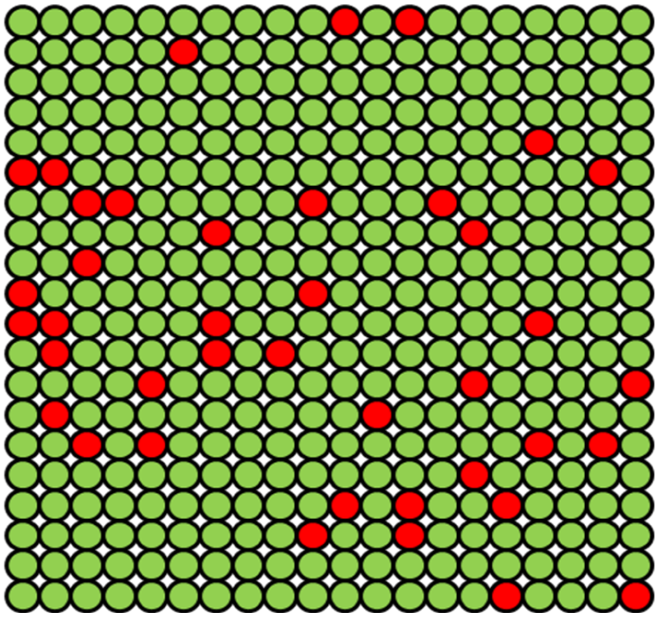


Abb. 2.12-3: Theoretische Population (n = 400) mit der Prävalenz eines qualitativen Attributs von 0,1 (10 %). Die Verteilung des Attributs wurde anhand von Zufallszahlen ermittelt.

Wiederum anhand von Zufallszahlen wurden jeweils 5 „Stichproben“ der Größe 20 und 30 gezogen. Die Ergebnisse sind in Tabelle 2.12-3 aufgelistet.

Tab. 2.12-3: Ergebnis von je 5 zufälligen „Stichproben“ der Größe 20 und 30 aus der „Population“ aus Abbildung 2.13-3, gereiht nach dem Anteil der „Treffer“. (Die wahre Prävalenz beträgt 0,1)

Größe	Anzahl (Anteil) „Treffer“	95 % Vertrauensbereich	99 % Vertrauensbereich
20	1 (0,05)	0,001-0,249	0,00-0,31703
	1		
	2 (0,1)	0,012-0,317	0,005-0,387
	3 (0,15)	0,032-0,379	0,0176-0,449
	5 (0,25)	0,087-0,491	0,058-0,560
30	1 (0,033)	0,0008-0,172	0,0002-0,223
	1		
	4 (0,133)	0,0376-0,307	0,0233-0,363
	5 (0,167)	0,0564-0,347	0,0368-0,404
	6 (0,2)	0,0771-0,386	0,0544-0,443

2.12.3 Interpretation von Nullzählern

Ein pharmazeutisches Unternehmen möchte ein Präparat für die Behandlung von Rindern auf den Markt bringen. In Vorversuchen mit dem Präparat an Mäusen haben einige der trächtigen Mäuse abortiert. Das Unternehmen lässt daher das Präparat an 10 trächtige Färsen testen.

Keine davon verwirft. Kann die Zulassungsbehörde nun davon ausgehen, dass die Anwendung dieses Präparates bei trächtigen Rindern sicher ist?

Hierzu einige Überlegungen anhand einer Grafik (Abb. 2.12-4):

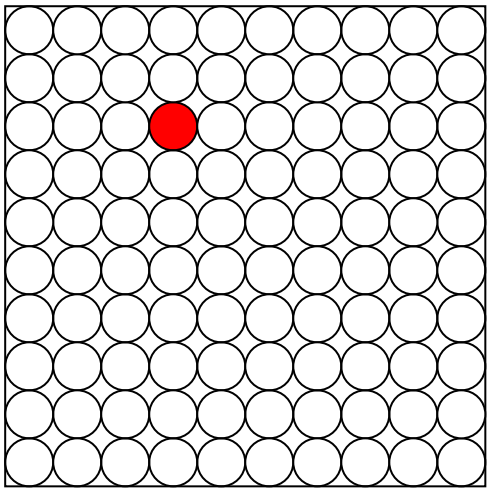


Abb. 2.12-4: Schematische Darstellung eines seltenen Ereignisses (Wahrscheinlichkeit 0,01 = 1 %)

In Abbildung 2.13-4 sind 100 Kreise, welche trächtige Rinder repräsentieren sollen, in 10 Reihen à 10 angeordnet. Der eine rote Kreis symbolisiert ein Individuum, bei welchem das fragliche Ereignis, hier also Abort nach Applikation des Mittels eintreten würde. In dieser Situation wären wir kaum überrascht, wenn in einer Stichprobe von 10 Tieren das Ereignis nicht beobachtet worden wäre.

Ähnlich wie bei dem Problem mit den BVDV-PI-Kälbern betrachten wir die Wahrscheinlichkeit, dass ein Abort nicht auftritt. Sie beträgt in der dargestellten Situation offensichtlich 0,99 und in einer Stichprobe der Größe 10 daher $0,99^{10} = 0,90$. Die komplementäre Wahrscheinlichkeit, also diejenige, dass in einer Stichprobe vom Umfang 10 mindestens ein Abort auftritt, ist 10 %. Mithin ist im Mittel nur in jeder 10. Stichprobe dieser Größe das Ereignis zu erwarten.

Die Frage, die es zu entscheiden gilt, ist, mit welcher maximalen Wahrscheinlichkeit (P_{max}) für einen Abort bei trächtigen Rindern das ermittelte Ergebnis von 0/10 vereinbar ist. Über die akzeptierte Irrtumswahrscheinlichkeit (üblicherweise entweder 5 % oder 1 %) muss entschieden werden.
Wiederum gehen wir von der Wahrscheinlichkeit aus, dass ein Abort nicht auftritt. Sie ist bei jedem Individuum $1 - P_{max}$, bei 10 Individuen daher $(1- P_{max})^{10}$. Die Wahrscheinlichkeit, mit der eine solche Serie auftritt, entspricht der zu wählenden Irrtumswahrscheinlichkeit, also beispielsweise 0,05 (5 %). Daher gilt:

$$\mathbf{(1- P_{max})^{10} = 0,05} \qquad \mathbf{(14)}$$

$1- P_{max} = 0,05^{1/10}$

$1- P_{max} = 0,74$

$P_{max} = 0,26$

Auf der Basis der beobachteten Serie von 0/10 ist das Risiko für Aborte nach Verabreichung des Präparats an Rinder mit 95 %iger Sicherheit maximal 26 %.

Es gibt eine Faustformel zu einfacheren Berechnung dieses maximalen Risikos. Sie lautet 3/n für eine Irrtumswahrscheinlichkeit von 5 %, 4,5/n für eine Irrtumswahrscheinlichkeit von 1 % und 6,9/n für eine Irrtumswahrscheinlichkeit von 0,1 %. Beispiel: Wird in einer Serie von 35 Nierenbiopsien bei Kühen keine nennenswerte Komplikation festgestellt, so liegt mit einer Sicherheit von99 % das Risiko für eine Komplikation bei maximal 4,5/35 = 0,129, also 12,9 % (berechneter Wert: 12,3 %). (Literatur: Hanley u. Lippman-Hand JAMA 249 [1983] 1743 – 1745).
Das so errechnete maximale Risiko darf nicht mit dem mittleren Risiko verwechselt werden! Wären in der genannten Serie von 35 Nierenbiopsien bei 3 Kühen Komplikationen aufgetreten, wäre das mittlere Risiko 8,6 % mit einem 95 % Vertrauensbereich (= Konfidenzintervall) von 1,8 bis 23,1 %.

2.12.4 Quantitative Merkmale:Abschätzung der mittleren Größe

Milchmastkälber werden durch Unterversorgung mit Eisen künstlich blutarm gehalten, damit ihr Fleisch „schön“ weiß ist, weil die offensichtlich unbelehrbaren Verbraucher das so wünschen. (Nebenbei bemerkt, geht diese Anämie mit Erhöhung der Infektionsanfälligkeit und damit Steigerung des Einsatzes von Antibiotika einher.) Die Tierschutz-Nutztierhaltungsverordnung vom 25. 10. 2001 schreibt in § 11 vor, dass bei Kälbern die Hämoglobinkonzentration im Blut im Mittel der Gruppe mindestens 6 mmol/L betragen muss. Eine Amtstierärztin soll einen Kälbermastbetrieb mit 2000 Kälbern kontrollieren und prüfen, ob der Betriebsleiter die Forderung der Verordnung einhält. So, wie der Verordnungstext lautet, ist eine Überprüfung nicht möglich, wenn nicht alle Tiere untersucht werden. Die Größe der „Gruppe“ ist nicht definiert. Offensichtlich ist es nicht praktikabel, bei allen 2000 Kälbern eine Blutprobe zu nehmen und erst recht nicht wiederholt.

Ein Vorschlag zur Formulierung der Vorschrift lautet: Es muss eine repräsentative Stichprobe von mindestens 30 Kälbern untersucht werden, und die Untergrenze des 95 %igen Vertrauensbereichs des Mittelwerts darf 6,0 mmol/L nicht unterschreiten.

95 %-Vertrauensbereich des Mittelwertes = : $\pm$ 1,96*$\sigma/\sqrt{n}$ (15)

Wenn aus früheren Untersuchungen bekannt ist, dass die Werte annähernd eine GAUSS-Verteilung zeigen und die Standardabweichung 0,3 mmol/L beträgt, dann müsste der errechnete Mittelwert mindestens 6,11 mmol/L betragen, wenn zu 95 % sichergestellt sein soll, dass der wahre Mittelwert nicht unter 6,0 mmol/L liegt ($1,96*0,3/\sqrt{30} = 0,107$).

2.12.5 Quantitative Merkmale: Abschätzung des Anteils über/unter einem Grenzwert

In dem Beispiel mit der Hämoglobinkonzentration bei Mastkälbern könnte man auch fragen, bei welchem Anteil der Gruppe die Hb-Konzentration unter dem Grenzwert von 6 mmol/L liegt, denn dies ist für jedes Einzeltier relevant. Unter der Voraussetzung, dass die Werte in etwa normal verteilt sind, kann dieser Anteil anhand von Mittelwert und Standardabweichung

bestimmt werden. [Die Funktion in EXCEL lautet normvert(cutoff;mittelwert;standardabweichung;wahr)]. Wenn z.B. der Mittelwert einer Stichprobe mit 6,2 mmol/L und die Standardabweichung mit 0,1 mmol/L errechnet worden sind, ergibt sich ein Anteil von etwa 2,3 % mit Werten unter 6,0 mmol/L. Schon bei einer etwas größeren Standardabweichung von 0,15 mmol/L steigt der Anteil auf über 9 %.

In Tabelle 2.12-4 sind die Anteile der Werte unter 6,0 mmol/l in Abhängigkeit von Mittelwert und Standardabweichung einer Stichprobe als Dezimalbrüche angegeben. Aufgeführt werden nur Werte über 0,01 (1 %).

Tab. 2.12-4: Zu erwartender Anteil der Werte unter 6,0 mmol/L in einer hinreichend normalverteilten Population von Milchmastkälbern, in Abhängigkeit von Mittelwert und Standardabweichung einer Stichprobe.

Standard-abwei-chung	Mittelwerte 6,1	6,2	6,3	6,4	6,5	6,6
0,06	0,048					
0,08	0,106					
0,1	0,159	0,023				
0,12	0,202	0,048				
0,14	0,238	0,077	0,016			
0,16	0,266	0,106	0,030			
0,18	0,289	0,133	0,048	0,013		
0,2	0,309	0,159	0,067	0,023		
0,22	0,325	0,182	0,086	0,035	0,012	
0,24	0,338	0,202	0,106	0,048	0,019	
0,26	0,350	0,221	0,124	0,062	0,027	0,011
0,28	0,360	0,238	0,142	0,077	0,037	0,016
0,3	0,369	0,252	0,159	0,091	0,048	0,023

2.12.6 Quantitative Merkmale: Metabolische Profiltests

Auf der Basis der verlockenden Vorstellung, dass es sinnvoller und leichter ist, sich ankündigende Gesundheitsstörungen im „präklinischen“ Bereich zu erfassen und zu behandeln, wurden die so genannten metabolischen Profiltests entwickelt. Hierzu werden die Konzentrationen verschiedener klinisch-chemischer Parameter im Blut und/oder in der Milch bei eng definierten Gruppen (zum Beispiel Milchkühe in den ersten 14 Tagen der 2. Laktation) bestimmt. Die Referenzintervalle dieser Parameter sind dabei meist enger gefasst als bei der Untersuchung klinisch kranker Tiere, im Extremfall sind sie sogar betriebsspezifisch. Häufung von Werten außerhalb der Referenzintervalle wird als Indikation für gezielte Untersuchungen und für „Metaphylaxe“ (Intervention bei besonders krankheitsgefährdeten Tieren) angesehen, also z.B. für Änderung der Ration und/oder für die Verabreichung von Wirkstoffen.

Auch wenn nachgewiesen werden kann, dass beispielsweise Kühe, die in der Frühlaktation Labmagenverlagerung bekommen, schon einige Zeit davor Abweichungen gewisser Laborwerte zeigten, ist zum einen die Spezifität dieser Befunde fraglich, und zum anderen stehen nicht immer konkrete spezifische und absolut wirksame „metaphylaktische“ Maßnahmen zur Verfügung. Außerdem müssten diese Profiltests an allen Tieren im fraglichen Stadium durchgeführt werden, im Extremfall sogar mehrmals.

2.12.7 Optimale Auswahlstrategie bei regionalen Untersuchungen auf Seuchenfreiheit

Die Frage, wie begrenzte Mittel für regionale Untersuchungen auf Seuchenfreiheit optimal eingesetzt werden, ist nicht trivial. Es geht dabei um die Kombination von Anteil zu untersuchenden Herden und Anteil zu untersuchender Tiere in den Herden.
Einen Vorschlag dazu haben Cameron u. Baldock publiziert (Prev. Vet. Med. 34 (1998) 19-30).
Eines ist jedoch leicht einzusehen. Wenn es darum geht, möglichst sicher Hinweise auf die Anwesenheit der gesuchten Infektion (z. B. Paratuberkulose) zu finden, ist ein möglichst sensitiver Test angezeigt. Bei der Auswahl der Probanden müssen jeweils die ältesten Tiere für eine Untersuchung auf spezifische Antikörper herangezogen werden, weil sie innerhalb einer Herde die größten Chancen hatten, sich mit dem Erreger auseinanderzusetzen.

2.12.8 Outbreak Investigation

Einfache epidemiologische Prinzipien sind intuitiv erfassbar und werden in der tierärztlichen Praxis auch angewandt, ohne dass explizit von Epidemiologie die Rede ist.

Am Anfang steht die Definition eines Falles. Das ist keine ganz triviale Angelegenheit. Die Definition sollte nicht zu breit gefasst werden (z.B. „frisst weniger“), damit nicht zu viel „Beifang“ erfasst wird, andererseits aber nicht zu eng (z.B. „fieberhafte Dyspnoe mit Verdichtungen der Lunge über eine Höhe von > 50 % der maximalen Lungenhöhe“), so dass die Erfüllung nur durch umfangreiche Untersuchungen des Einzeltiers geklärt werden kann. Ein akzeptabler Mittelweg wäre etwa „Atemwegserkrankung“.

Nützlich ist eine retrospektive Auflistung der Ereignisse, nicht nur der neuen Fälle, sondern auch z.B. Anbruch einer neuen Silage-Charge von Tag zu Tag. Die Dokumentation der neuen Fälle ergibt eine epidemiologische Kurve, deren Aussehen Hinweise auf die Art der Ursache geben kann.

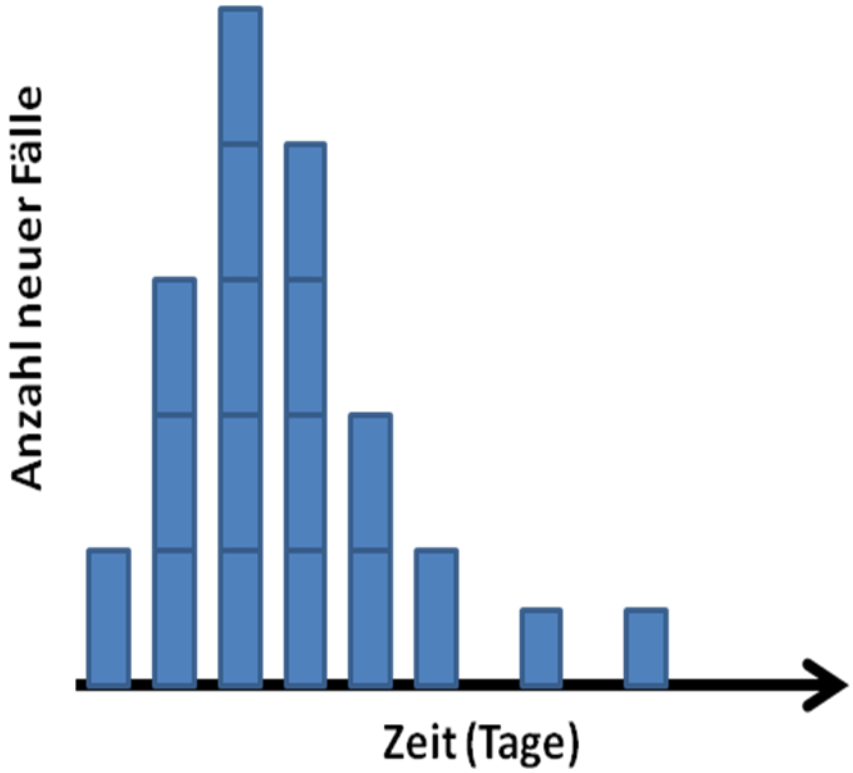


Abb. 2.12-5: Epidemiologische Kurve, die für eine Punkt-Exposition spricht

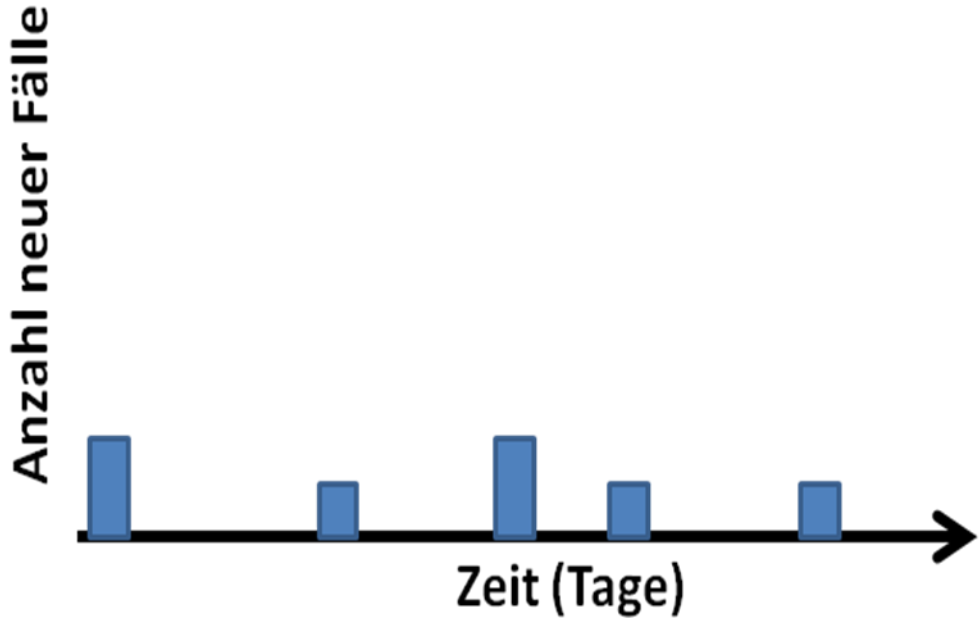


Abb. 2.12-6: Epidemiologische Kurve, die für eine anhaltende Exposition spricht

Das Vorgehen bei dem Versuch, die Ursache(n) einer auffälligen Häufung von Krankheits- und/oder Todesfällen in einer Population zu finden, entspricht zunächst einmal dem gesunden Menschenverstand. Es sollte in einer Teilgruppe, die gegenüber einer vermuteten Ursache

exponiert war oder ist, der relative Anteil der Fälle höher sein als in nicht exponierten Teilgruppen. Es gibt verschiedene Maßzahlen zur Prüfung auf einen möglichen Zusammenhang zwischen einer vermuteten Ursache und dem Auftreten von Fällen, z. B.:

- Attributable Risk
- Relatives Risiko
- Odds Ratio

Tab. 2.12-5: Vierfeldertafel zur Darstellung von Attributable Risk, Relativem Risiko und Odds Ratio

	Krankheit vorhanden	**Krankheit nicht vorhanden**
Exponiert	**a**	**b**
Nicht exponiert	**c**	**d**
	a + c	**b + d**

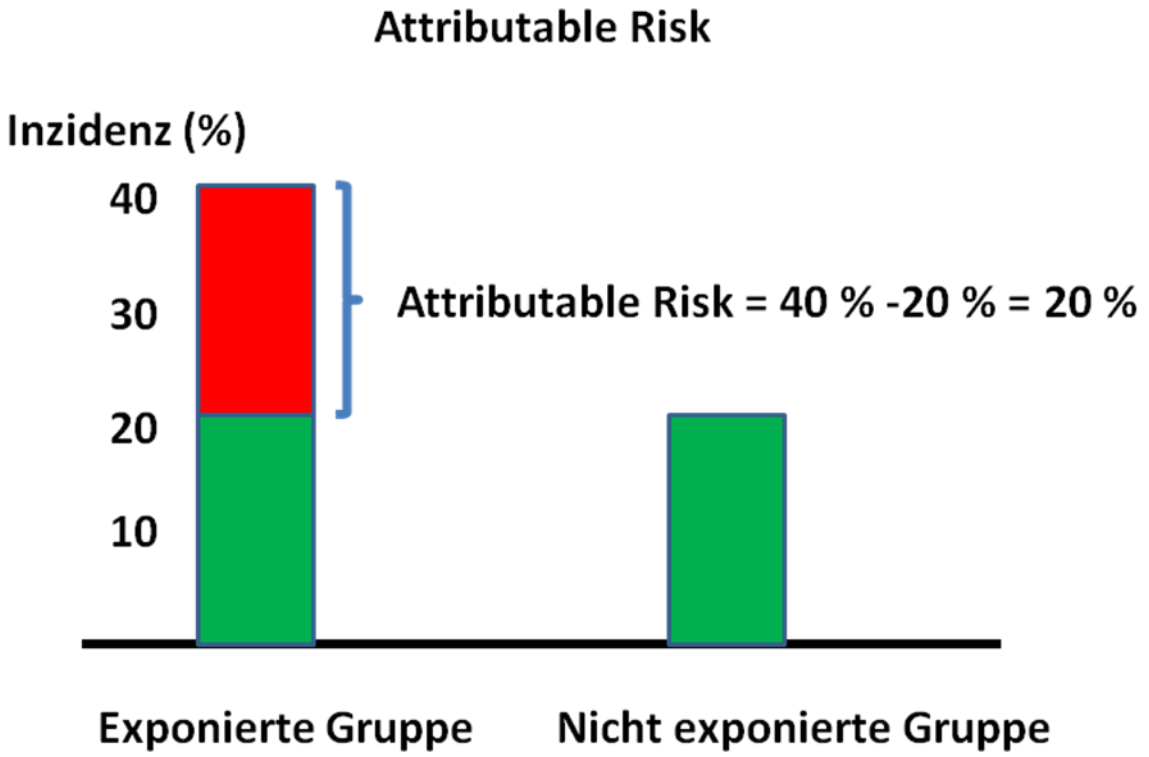


Abb. 2.12-7: Attributable Risk

Attributable Risk:Im gezeigten Beispiel ist der Anteil von Fällen in der gegenüber einem bestimmten Faktor exponierten Teilgruppe 40 %, in der nicht exponierten Gruppe dagegen nur 20 %. Das wird so interpretiert, dass 20 %-Punkte der im fraglichen Zeitraum aufgetretenen Fälle in der exponierten Gruppe dem Faktor zugeschrieben werden können. In den Angaben aus Tabelle 2-12.5 entspricht Attributable Risk (a/(a+b)) – (c/(c+d)). Allerdings wird was Risiko dann als Dezimalbruch errechnet. Ist der errechnete Wert negativ, hat die Exposition gegenüber dem Faktor eher eine schützende Wirkung.
Attributable Risk (AR) gibt auch einen Hinweis darauf, welcher Anteil von Fällen durch Ausschaltung der Exposition zu verhindern wäre.

Relatives Risiko (RR) ist das Verhältnis der Risiken, zum Fall zu werden, in der exponierten Gruppe zu der in der nicht exponierten Gruppe. Als Formel: RR = (a/(a+b))/(c/(c+d)). Auch hier ist das Ergebnis ein Dezimalbruch. Ein angenommener Wert von 1,3 bedeutet, dass das Risiko, Fall zu sein, in der exponierten Gruppe um 30 % höher als in der nicht exponierten Gruppe ist. Hier bedeuten Werte < 1 schützende Wirkung.
In den Medien werden mitunter Zahlen zu relativen Risiken ohne Angaben der zugrunde liegenden absoluten Zahlen genannt. Wenn in einer exponierten Gruppe der Größe 1.000.000 in einem bestimmten Zeitraum 5 Fälle auftreten, in einer nicht exponierten Gruppe der Größe 800.000 dagegen 2 Fälle, ist das relative Risiko aufgrund der Exposition 2, also um 100 % höher. Das mag spektakulär klingen, soll es vermutlich auch, aber das absolute Risiko ist in der exponierten Gruppe mit 1:200.000 nicht gerade riesig.

Odds Ratio (s. 4.4)

3. Therapie

3.1. Beurteilung der Wirksamkeit von Interventionen

Bei Angaben über Behandlungserfolge sind unter anderen folgende Aspekte zu beachten:

- Falldefinition (Goldstandard)
- Schweregrade/Stadien
- Prüfung auf „confounders“ (z.B. anderweitige Behandlung oder Erkrankung, Umweltfaktoren)
- Stichprobengröße
- Kontrollgruppe (unbehandelt oder mit bisheriger Therapie behandelt)
- Methode der Zuteilung zu Behandlungs- und Kontrollgruppe(n)
- Wie oft wurde kontrolliert?
- Kriterien für Besserung und Heilung (Zielgröße, „outcome“)?
- Statistisch signifikanter vs. klinisch relevanter Unterschied
- Wer hat behandelt und wer hat Befunde erhoben? („Maskierung“, „Blinding“)
- Protokolle (Good Clinical Practice) (http://www.gesetze-im-internet.de/gcp-v/index.html)
- Good epidemiological practice: http://webcast.hrsa.gov/conferences/mchb/mchepi_2009/communicating_research/Ethical_guidelines/IEA_guidelines.pdf

Ein Therapieeffekt ist umso leichter „mit bloßem Auge“ erkennbar,

- je einheitlicher der Krankheitsverlauf ohne Behandlung ist
- je früher die Wirkung nach der Behandlung eintritt
- je drastischer er ist
- je regelmäßiger er eintritt

Neben statistischen Tests, zum Beispiel Chi-Quadrat-Test für den Vergleich der Proportionen qualitativ definierter Kriterien (also z.B. Heilung – wie auch immer definiert) oder t-Test für der Vergleich quantitativ definierter Kriterien (also z.B. absolute Abnahme der Körpertemperatur 24 h nach Applikation eines Antiinfektivums bei Jungrindern mit Pneumonie) zwischen Behandlungsgruppen, gibt es noch weitere Maße, zum Beispiel die Anzahl nötiger Patienten (NNT = Numbers needed to treat) zur Reduktion der Anzahl ungünstiger Ausgänge um eins. Angenommen, beim Vergleich der bisherigen Therapie bei Krankheit K mit einer neuen Behandlungsmethode würden 15 der 100 Patienten aus der Kontrollgruppe (bisherige Therapie) und 5 der 100 Patienten aus der Versuchsgruppe (neue Therapie) sterben. Der Unterschied ist im Chi-Quadrat-Test signifikant. Die Reduktion der Letalität beträgt 10/15, also 2/3, was recht eindrucksvoll ist. Man kann sich aber auch fragen, wie viele Patienten mit der neuen Methode behandelt werden müssen, damit ein Todesfall verhindert wird.
Dazu ist zunächst die absolute Reduktion des Risikos (hier Letalität) zu berechnen: ARR = Anteil des Ereignisses in der Kontrollgruppe – Anteil des Ereignisses in der Versuchsgruppe = 0,15 – 0,05 = 0,1.

Anzahl nötiger Patienten (NNT) = 1/ARR = 1/0,1 = 10 **(16)**
Mit einem 95 % Vertrauensintervall von 1,6 bis 11,3.

3.2 Hypothesen-basierte Forschung

Das Prinzip der hypothesen-basierten Forschung geht auf Sir Karl Popper zurück. Auf der Basis der Erkenntnis, dass man die Wahrheit einer Annahme nie direkt mit absoluter Sicherheit feststellen kann, postulierte er, dass man durch wiederholtes Scheitern von Versuchen, sie zu falsifizieren, immer sicherer werden kann, dass sie nicht falsch ist.

Dieses Prinzip lässt sich anhand einer Scherzfrage illustrieren: Wie macht man eine Statuette eines Elefanten? Antwort: Man nimmt einen Steinblock uns schlägt alles weg, was nicht wie ein Elefant aussieht. Was stehen bleibt, muss dann einem Elefanten immer ähnlicher werden.

Wenn geprüft werden soll, ob zwischen den Resultaten zweier Behandlungsalternativen ein Unterschied ist, lautet die zu falsifizierende Nullhypothese, dass kein Unterschied besteht, die Datensätze der beiden Versuchsgruppen (Verum vs. Placebo oder Versuchspräparat vs. Standardbehandlung) sozusagen aus demselben Topf stammen, dessen Daten durch Lage des Mittelwerts und die Größe der Streuung gekennzeichnet sind.

Dass kein Unterschied besteht, kann nicht nachgewiesen werden, sondern allenfalls mit einem vorgegebenen Grad an Sicherheit, dass ein Unterschied, wenn er denn existiert, eine bestimmte Größe nicht überschreitet. Die Größe eines nachweisbaren Unterschieds hängt von der Streuung der Daten, dem Grad an gewünschter Sicherheit (also der akzeptierten Irrtumswahrscheinlichkeit) und der Stichprobengröße ab. Diese vier Größen hängen also zusammen, und man kann jeweils eine mithilfe der anderen berechnen. Sinnvoller Weise wird vor einem Versuch die nötige Stichprobengröße berechnet. Wird das nicht gemacht, und die Untersuchung ergibt keinen statistisch signifikanten Unterschied, kann retrospektiv mit der so genannten Poweranalyse berechnet werden, ab welcher Größe ein tatsächlich bestehender Unterschied hätte nachgewiesen werden können. Aus dem Zusammenhang der vier Größen wird auch klar, dass statistische Signifikanz (= Größe der Irrtumswahrscheinlichkeit) keine fixe Größe ist, sondern eine Funktion der übrigen drei Größen ist.
Es gibt zwei Arten von Irrtümern (oder Fehler), die sich am besten anhand eines Vergleichs zwischen den Daten einer einer Versuchs- und einer Kontrollgruppe beschreiben lassen. Der Fehler erster Art (α) besteht darin, einen Unterschied anzunehmen, der in Wirklichkeit nicht existiert („false alarm“). Der Fehler zweiter Art (ß) besteht darin, einen tatsächlich bestehenden Unterschied zu übersehen („miss“). Welcher Fehler klein gehalten werden soll, hängt von der fachlichen Bedeutung der beiden Fehler ab. Wenn eine neue Therapie mit (vermuteter) Verbesserung der Heilungsaussichten aber gravierenden Nebenwirkungen mit einer etablierten Therapie verglichen werden soll, muss der Fehler erster Art (also die fälschliche Annahme einer signifikanten Verbesserung der Heilungsaussichten) besonders klein gehalten werden. Wenn eine neue Therapie für eine bisher unheilbare, schwerwiegende Krankheit getestet werden soll, muss der Fehler zweiter Art (also das Übersehen einer tatsächlich vorhandenen Verbesserung der Heilungsaussicht) klein gehalten werden.
Die Festlegung von $\alpha = 0{,}05$ ist eine Konvention und steht nicht in der Bibel. Sie wird meist aus Bequemlichkeit übernommen. Besser wäre die fachlich begründete Wahl der beiden Fehlerarten sowie die Angabe der retrospektiv ermittelten Größen.

3.3 Klinische Entscheidungsanalyse

In der Medizin müssen laufend Entscheidungen gefällt werden, ohne dass alle wünschenswerten Informationen zur Verfügung stehen. Es handelt sich also um Entscheidungen unter Unsicherheit. Diese Situation betrifft natürlich nicht allein die Medizin sondern viele andere Bereiche des Lebens auch. Es gibt mehrere Möglichkeiten, Entscheidungen dieser Art rationaler zu gestalten. Eine davon ist die so genannte Entscheidungstheorie oder Entscheidungsanalyse (EA).

Die klinische Entscheidungsanalyse (engl.: Clinical decision analysis) ist ein explizites, quantitatives und präskriptives Verfahren, das zur Unterstützung und Erweiterung, nicht aber als Ersatz der klinischen Beurteilung einer Situation dienen soll.

Es ist explizit, weil es den Kliniker zwingt, ein Problem in seine wichtigsten Komponenten zu zerlegen, diese zu benennen, sie einzeln zu untersuchen und dann wieder zusammenzusetzen.

Es ist quantitativ, weil es den Kliniker zwingt, möglichen Ereignisse bestimmte, also quantifizierte Wahrscheinlichkeiten und möglichen Ergebnissen bestimmte (quantitative) Werte zuzuordnen.

Es ist präskriptiv, weil es eine Strategie als die unter den gegebenen oder angenommenen Umständen beste vorschreibt.

Es ist also nicht deskriptiv, hat nicht zum Zweck zu beschreiben, wie Kliniker tatsächlich zu Entscheidungen gelangen.

Die EA besteht aus fünf grundlegenden Schritten:

1. Identifikation und Begrenzung des Entscheidungsproblems. Je präziser die Beschreibung, umso relevanter sind die Ergebnisse, die sich mit Hilfe der EA erzielen lassen. Also z.B. nicht nur „Kuh mit Durchfall“, sondern z.B. „einzelne erwachsene Kuh mit anhaltendem Durchfall und erhaltener Fresslust“.

2. Strukturierung des Entscheidungsproblems unter Beachtung der logischen und zeitlichen Reihenfolge; Konstruktion eines Entscheidungsbaum

3. Charakterisierung der zur Ausfüllung des Entscheidungsbaums benötigten Informationen

4. Durchführung der Entscheidungsanalyse ⇨Auswahl der bevorzugten Handlungsweise

5. Überprüfung der Entscheidung auf Stabilität (Sensitivitätsanalyse und Schwellenanalyse)

Zu den einzelnen Punkten:

Der Ausgangspunkt ist von großer Bedeutung. Je genauer er definiert ist, desto relevanter und realistischer sind die Ergebnisse.

Die an dieser Stelle möglichen Handlungen werden als Zweige eines so genannten Entscheidungsknotens (engl.: decision node) dargestellt. Entscheidungsknoten werden nach Übereinkunft durch kleine Quadrate symbolisiert. Ein für viele Situationen in der klinischen Praxis passender Entscheidungsknoten könnte etwa so aussehen, wie in Abb. 3.2-1 dargestellt.

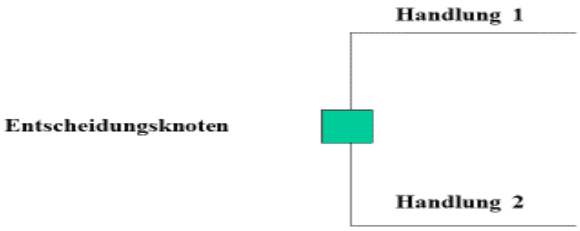


Abb. 3.3-1: Entscheidungsknoten in der Entscheidungsanalyse

Die relevanten Ereignisse mit probabilistischem Charakter werden als Abzweigungen von einem so genannten Zufallsknoten (engl.: chance node) dargestellt. Zufallsknoten werden durch Kreise symbolisiert (s. Abb. 3.2-2).

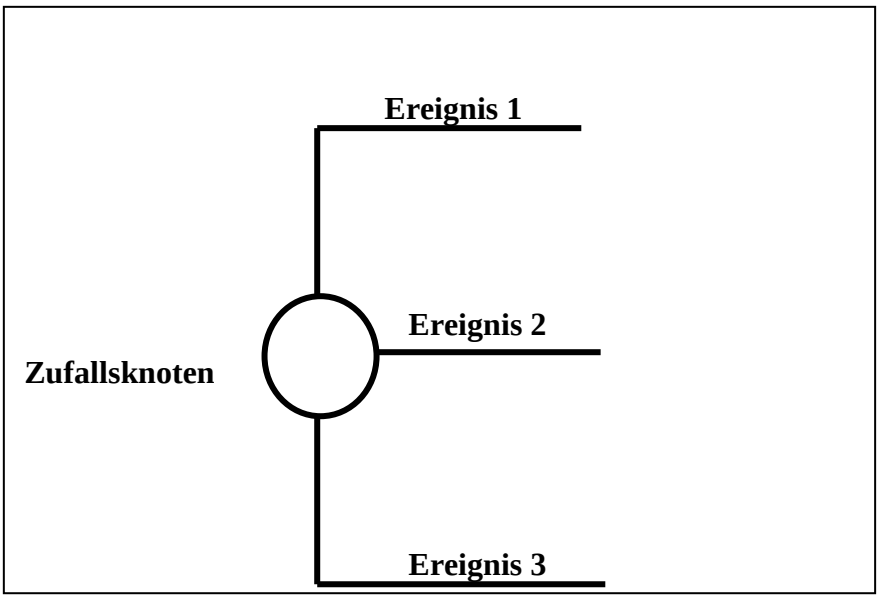


Abb. 3.3-2: Zufallsknoten in der Entscheidungsanalyse

Es muss sich nicht immer um Zufallsereignisse handeln, was auch immer darunter zu verstehen sein mag. Wichtig ist nur, dass sie durch denjenigen, der die Entscheidungsanalyse durchführt, vor ihrem Eintritt nicht kontrollierbar, wohl aber nach ihrem Eintritt erkennbar und unterscheidbar sind.

Die an einem Zufallsknoten möglichen Ereignisse müssen sich gegenseitig ausschließen, zusammen aber alle relevanten Möglichkeiten einschließen. Mit anderen Worten: eines der vorgesehenen Ereignisse muss eintreten, und nur eines darf eintreten. Das wiederum bedeutet, dass die Summe der Wahrscheinlichkeiten der einzelnen Ereignisse an einem Zufallsknoten als Dezimalbrüche immer 1,0 betragen muss (s. Abb. 3.2-3).

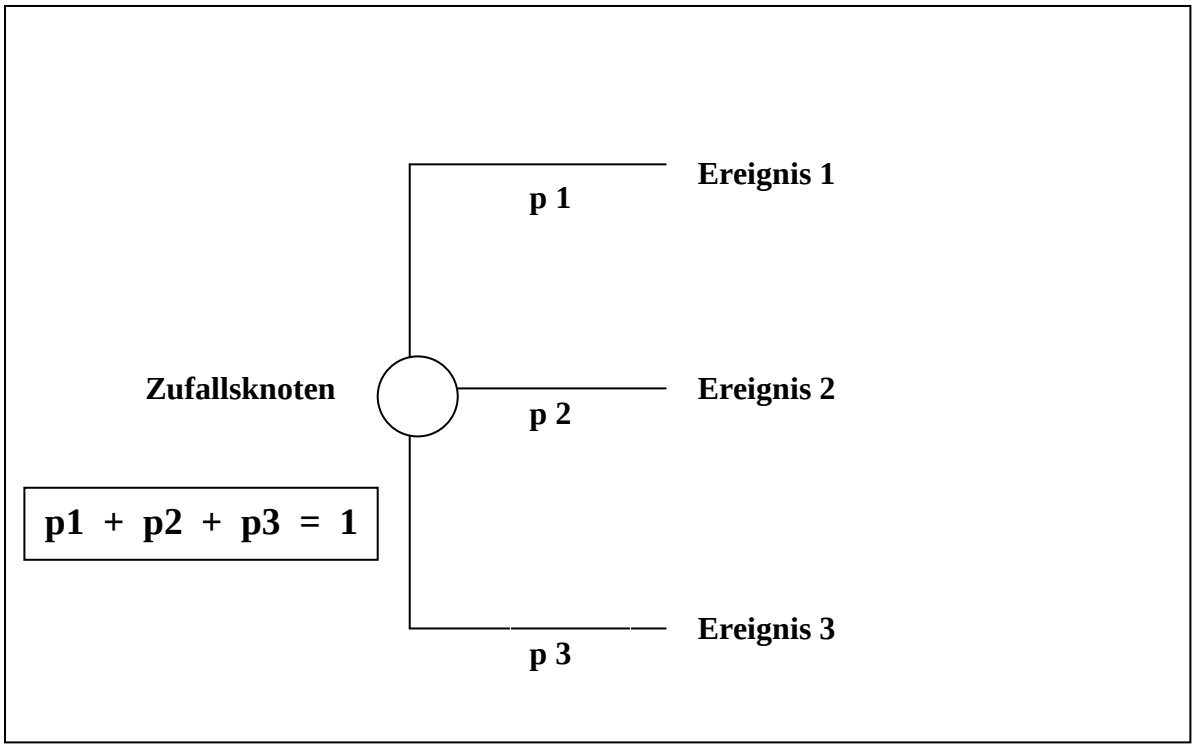


Abb. 3.3-3: Wahrscheinlichkeiten der einzelnen Ereignisse an einem Zufallsknoten

Man darf sich das nicht als wunderbare Fügung des Schicksals oder als Naturgesetz vorstellen, sondern diese Situation entsteht durch korrekte Überlegungen und zweckmäßige Definitionen seitens des Entscheidungsanalytikers. Sind beispielsweise nur zwei Ereignisse von Interesse, aber weitere möglich, können deren Wahrscheinlichkeiten summarisch als 1–(p1 + p2) zusammengefasst werden.

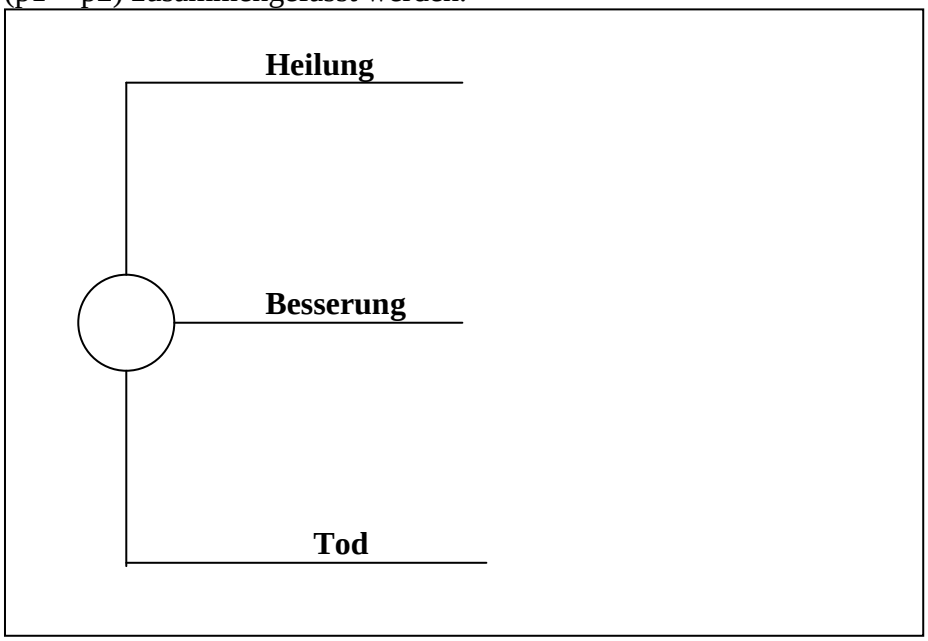


Abb. 3.3-4: Beispiel für Zufallsknoten

Frage 3.3-1: Ist der Zufallsknoten in Abb. 3.2-4 optimal konstruiert?

Bei der Abschätzung der Wahrscheinlichkeit der einzelnen Ereignisse zeigen sich die Vorteile der Entscheidungsanalyse recht deutlich, denn die Bestandteile des Problems können einzeln sozusagen unter die Lupe genommen und getrennt so gut wie möglich bearbeitet werden. Dabei zeigt sich auch, wo gegebenenfalls wesentliche Informationslücken bestehen. Diese Lücken bestehen natürlich nicht nur bei der formalisierten Bearbeitung des betreffenden Problems, sondern erst recht bei der traditionellen, informellen, mehr oder weniger intuitiven "Bearbeitungsweise".

Mögliche Quellen für die benötigten Informationen sind einschlägige Literatur (insbesondere Meta-Analysen), lokale Datenbanken und Expertenbefragung.

Frage 3.3-2: Sie möchten einen netten Abend in einem kleinen Künstlerzirkus verbringen. An der Kasse werden Sie mit folgender Wahl konfrontiert: Zahlen Sie Euro 20,-- Eintritt oder würfeln Sie?

Bei einer gewürfelten 1 bekommen Sie Euro 30,-- ausbezahlt,
bei einer 2 oder 3 bekommen Sie Euro 2,-- ausbezahlt,
bei einer 4, 5 oder 6 müssen Sie Euro 50,-- bezahlen.

Wie entscheiden Sie sich?

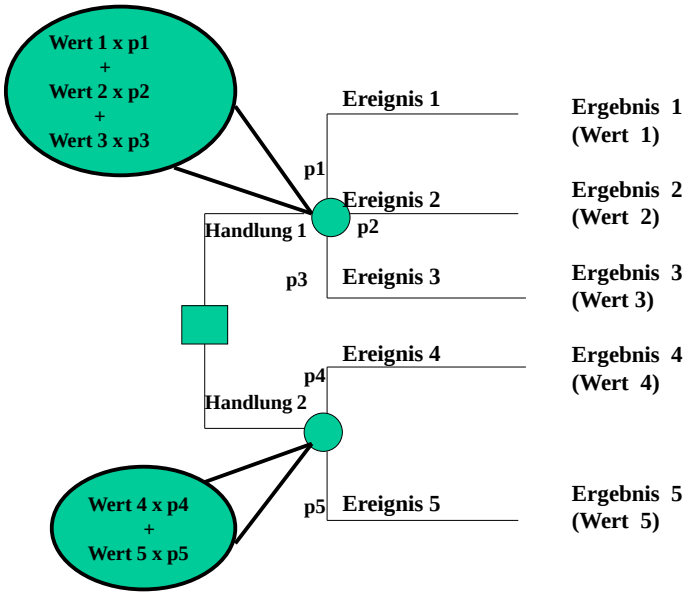


Abb. 3.3-5: Fertiger Entscheidungsbaum

Die Bewertung der möglichen Ergebnisse (also z. B. der Nutzwert einer Kuh nach Eintritt des Ereignisses „Heilung“ abzüglich der damit verbundenen Kosten) ist von größter Bedeutung. Es können jedoch erhebliche Schwierigkeiten damit verbunden sein. Verschiedene Personen können nämlich ein bestimmtes Ereignis ganz verschieden bewerten. Daher muss die Bewertung innerhalb einer EA strikt aus der Sicht einer Person vorgenommen werden. Diese Schwierigkeiten sind in der Humanmedizin sicher schwerwiegender als in der Nutztiermedizin, wo die wesentlichen Faktoren auf der ökonomischen Ebene liegen. In der Humanmedizin müssen dagegen mitunter inkommensurable Größen wie Behandlungskosten, Lebensdauer und Lebensqualität miteinander verglichen werden.

Im vierten Schritt, der Auswahl einer Handlungsweise, erfolgt eine Synthese aller verfügbaren Informationen. Dazu werden alle Ergebnisse mit der Wahrscheinlichkeit ihres Eintretens multipliziert und diese Produkte an jedem Zufallsknoten addiert. Das kann über mehrere Schritte erfolgen, bis jeder der am Ausgangspunkt möglichen Handlungen ein mittlerer zu erwartender Wert zugemessen werden kann. In Abb. 3.2-5 ist das Prinzip anhand eines einfachen Entscheidungsbaums dargestellt. Es werden, von rechts beginnend, die Werte der Ergebnisse am Ende er einzelnen Äste eines Zufallsknotens mit ihren zugehörigen Wahrscheinlichkeiten multipliziert und die Produkte an jedem Zufallsknoten addiert, was den mittleren zu erwartenden Wert eines jeden Zufallsknotens ergibt. Diese Prozedur wird nach links fortschreitend fortgesetzt, bis jede der am initialen Entscheidungsknoten möglichen Handlungen bewertet werden kann. Dabei können „unterwegs“ schon Äste mit offensichtlich ungünstigem Ergebnis „gekappt“ werden.

Oft kann nicht für jede relevante Information ein genauer Wert mit Sicherheit angegeben werden, wohl aber ein plausibler Bereich. Im letzten Schritt, der Sensitivitätsanalyse, wird daher geprüft, wie stabil oder empfindlich das Ergebnis der Entscheidungsanalyse ist, wenn die zugrunde gelegten Annahmen über die Wahrscheinlichkeiten von Ereignissen oder die Bewertung von Ergebnissen in einem plausiblen Rahmen verändert werden. Im einfachsten Fall wird jeweils eine Annahme isoliert verändert, während alle anderen konstant gehalten werden. Wenn sich dabei eine Strategie stets als besser erweist als eine andere, sind auch scheinbar kleine Unterschieden im zu erwartenden Nutzen oder Verlust von Bedeutung. Die Bedeutung einer Annahme ist dann als kritisch anzusehen, wenn es innerhalb ihres plausiblen Rahmens einen Schwellenwert gibt, bei dessen Über- oder Unterschreitung das Ergebnis „kippt“.
Es sind auch Mehr-Weg-Sensitivitätsanalysen möglich. Ihre Ergebnisse werden meist graphisch dargestellt.

Warum Entscheidungsanalyse?
Zunächst hat schon das verwendete Vokabular eine gewisse Bedeutung. Es sind quantitative Angaben und bewusste Bewertungen erforderlich. Dies kann die Kommunikation über das anstehende Problem erleichtern.

Die EA kann zur Klärung von Meinungsverschiedenheiten beitragen. So kann relativ leicht festgestellt werden, wo die Ursache für Meinungsverschiedenheiten lokalisiert ist (Struktur des Baums, Wahrscheinlichkeiten, Werteinschätzung) und welche Informationen geeignet sind, die Meinungsverschiedenheiten zu beseitigen.

EA lehrt, zwischen Entscheidungen und Ergebnissen dieser Entscheidung zu differenzieren.

Eine auf der Basis der verfügbaren Information richtige Entscheidung kann zu einem unerwünschten Ergebnis führen und umgekehrt. Der Eintritt eines günstigen Ergebnisses ist kein Beweis für die Richtigkeit der vorausgegangenen Entscheidung.

EA ist vom Prinzip her korrekt. Es gibt keine theoretischen Einwände dagegen. Dass in der Realität oft auch nicht-rationale Aspekte Entscheidungen beeinflussen können, ändert nichts an der prinzipiellen Korrektheit der EA.

EA lehrt, die Informationen zu erkennen und zu suchen, die für Entscheidungen notwendig sind.

EA lehrt zu erkennen, dass Informationen nur dann von praktisch klinischem Wert sind, wenn sie Entscheidungen beeinflussen können. Natürlich kann es daneben auch noch andere Aspekte geben, z.B. wissenschaftliche oder forensische.

EA kann auch zur Beurteilung von Strategien zur Krankheitsbekämpfung auf regionaler, nationaler oder globaler Ebene im Sinne einer Kosten-Nutzen-Analyse eingesetzt werden. So hat eine solche Analyse dazu geführt, dass die bis 1991 durchgeführten jährlichen Impfungen gegen Maul- und Klauenseuche beendet wurden.

Es gibt kommerzielle Computerprogramme für die EA, aber für die Bearbeitung der meisten klinischen Probleme reichen Tabellenkalkulationsprogramme aus.

Eine klinisch relevante EA soll nun am Beispiel der Frage, ob in einem Bullenmaststall gegen enzootische Bronchopneumonie (EBP, „Rindergrippe“) geimpft werden soll, durchgespielt werden. Gleich zu Beginn muss betont werden, dass nur wirtschaftliche Aspekte berücksichtigt werden, also nicht zum Beispiel der Aspekt der möglichen Vermeidung von Pneumonie-bedingtem Leiden. Wenn diesem Aspekt überragende Bedeutung zugemessen wird, gibt es kein Entscheidungsproblem, solange die Überzeugung besteht, dass durch Impfung überhaupt eine Reduktion der Inzidenz von EBP zu erreichen ist, und durch die Impfung selbst kein oder nur vernachlässigbares Leiden verursacht wird.

In Abb. 3.2-7 ist der Entscheidungsbaum dargestellt. An seinem Beginn steht die Entscheidung für oder gegen Impfung. Danach hat sozusagen die Natur das Wort, denn entweder tritt behandlungsbedürftige EBP auf oder nicht. Da es bei EBP für den weiteren Verlauf von erheblicher Bedeutung ist, wann ein erkranktes Tier entdeckt wird, sind hier die drei Möglichkeiten frühe, späte und sehr späte Entdeckung vorgesehen. Hier kommen die Aufmerksamkeit und das „Auge“ des Personals zum Ausdruck. Es wird davon ausgegangen, dass alle als krank erkannten Rinder antibakteriell behandelt werden. Bei früher Entdeckung und unverzüglicher sachgerechter Behandlung sind die Heilungschancen deutlich höher und der Behandlungsaufwand niedriger als bei später und erst recht bei sehr später Entdeckung. Als jeweils mögliche Endpunkte werden Heilung chronische EBP mit deutlicher Leistungseinbuße und Tod angenommen.
Damit ist die Struktur des Entscheidungsbaums erstellt. Er hat 10 Zufallsknoten und 20 Endpunkte (a … t).
Jedem einzelnen Endpunkt muss unter Beachtung der bis zu seinem Auftreten aufgelaufenen Kosten ein Wert zugeordnet werden. Dabei muss man sich entscheiden, ob Wert oder allein Kosten berücksichtigt werden sollen. Am übersichtlichsten ist es in hier betrachteten Fall, allein die Kosten und Verluste aufzuführen, wobei allerding auch zum Beispiel die Differenz zwischen dem Wert eines gesund gebliebenen und dem eines chronisch kranken Tieres als

Verlust berücksichtig werden muss. Zur Verdeutlichung sollen einige Beispiele besprochen werden. Bei Endpunkt a sind die Impfkosten und die einfachen Behandlungskosten zu berücksichtigen, bei Endpunkt b die Impfkosten, erhöhte Behandlungskosten sowie die Differenz zwischen dem Erlös für ein (wieder) gesundes Tier und dem für chronisch krankes Tier. (Ob auch noch erhöhte Futterkosten für die Verlängerung der Mastperiode berücksichtigt werden sollen, ist „Geschmackssache". Man kann sie auch in der Differenz im Erlös „unterbringen".). Im Endpunkt c kommen zusätzlich noch die Kosten für Tötung hinzu. (Man könnte argumentieren, dass der Entscheidungsbaum hier noch insofern erweitert werden müsste, als Tod nach perakutem Verlauf und damit eventuell ohne oder nur mit geringen Behandlungskosten auch noch als Möglichkeit berücksichtigt werden müsste. Das ist kein prinzipielles Problem. Zur Demonstration wird der Entscheidungsbaum in seiner hier gezeigten Form aber als ausreichend angesehen.)
Endpunkt t verkörpert das für den Besitzer ideale Ergebnis, nämlich keine hier zu berücksichtigenden Kosten.

Nun muss jedem Ast eines Zufallsknoten eine Wahrscheinlichkeit zugeordnet werden, wobei zu beachten ist, dass sich die Wahrscheinlichkeiten für die Äste eines Zufallsknotens zu 1 addieren müssen. Die Wahrscheinlichkeit für den oberen Ast des Zufallsknotens 2 entspricht der zu erwartenden Inzidenz von EBP, die des unteren Astes daher dem Komplement zu 1,0. Im Unterschied zwischen den Wahrscheinlichkeiten der oberen Äste der beiden Zufallsknoten 1 und 2 kommt die erwartete Wirkung der Impfung zum Ausdruck.
Wie oben schon angedeutet, spiegeln die Wahrscheinlichkeiten für die Äste der beiden Zufallsknoten 3 und 4 die Qualität der Überwachung wider. Daraus ist ersichtlich, dass man die EA den Gegebenheiten eines bestimmten Betriebs anpassen kann und muss. Das gilt auch für die Frage, ob man die Wahrscheinlichkeiten an den sich entsprechenden Ästen jeweils als gleich annimmt oder davon ausgeht, dass geimpfte Tiere weniger genau überwacht werden.

Als Quelle für die benötigten Wahrscheinlichkeitsangaben kommen die einschlägige Literatur (wobei nach Möglichkeit unter dem Aspekt der „Evidenzbasierten Tiermedizin" erstellte Meta-Analysen verwendet werden sollten, die allerdings in der Tiermedizin noch selten sind), kritisch ausgewertete lokale Datenbasen und als Notbehelf „Expertenschätzungen" in Frage.

Die Wahrscheinlichkeiten für die sich entsprechenden Äste der Zufallsknoten 5, 6 und 7 unterscheiden sich in Abhängigkeit davon, am welchem der Äste von Zufallsknoten 3 sie abgehen. Analoges gilt für die Äste 8, 9 und 10.

Jedem der Äste an den Zufallsknoten 5 – 10 kann nun durch Multiplikation seines „Endwertes" mit der zugehörigen erwarteten Wahrscheinlichkeit ein mittlerer Wert zugeordnet werden. Diese Produkte werden an jedem Zufallsknoten dieser addiert. Damit ist ein mittlerer zu erwartender Wert für jeden der drei Äste der Zufallsknoten 3 und 4 ermittelt. Nun wird analog mit den Ästen diese beiden Knoten verfahren und noch einmal mit den Ästen der Zufallsknoten 1 und 2.

Als Resultat der EA ist damit ermittelt, welche der beiden Entscheidungen unter den angenommenen Kosten und Wahrscheinlichkeiten aus wirtschaftlicher Sicht die bessere ist.

Auf der Basis die hier gebotenen Erläuterungen ist es leicht möglich, die EA in einer Tabellenkalkulation durchzuspielen, was auch Spaß machen kann, wenn man gern mit Zahlen umgeht. Die dazu nötige Mathematik geht nicht über einfache Multiplikation und Addition hinaus.

Da viele der eingesetzten Daten Schätzwerte sind, ist eine Sensitivitätsanalyse unverzichtbar und in einer Tabellenkalkulation auch sehr leicht durchführbar. Das Ergebnis könnte verblüffen.

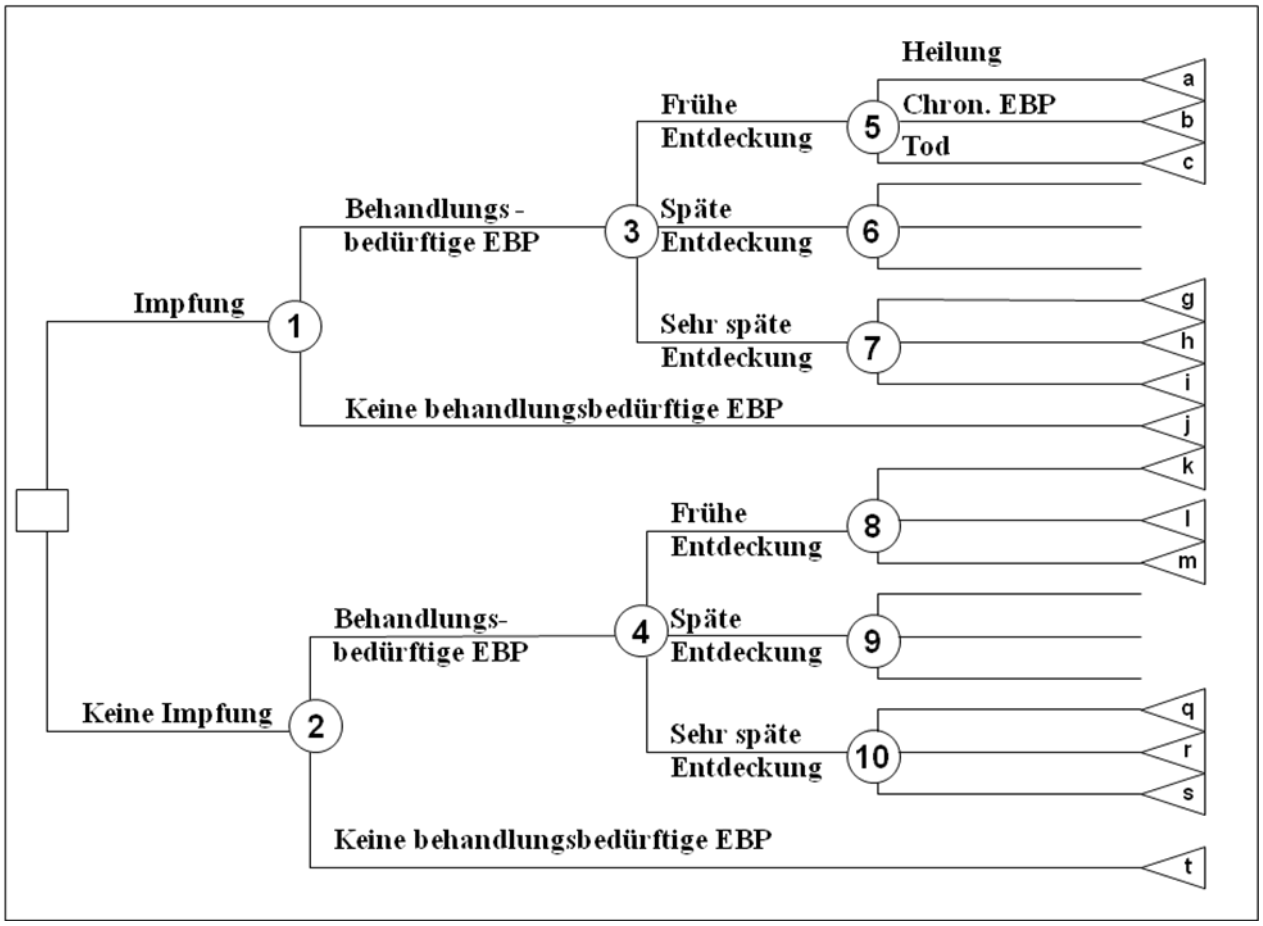


Abb. 3.3-7: Entscheidungsbaum für die Entscheidung für oder gegen Impfung gegen EBP (Erläuterungen im Text)

Frage 3.3.3: Pegbovigrastin ist ein Medikament, das bei Kühen nach Injektion die Inzidenz von klinischen Mastitiden in den ersten 30 Tagen der Laktation um durchschnittlich 25 % (95 % Vertrauensintervall 14 bis 35 %) reduziert. Der Einsatz pro Kuh kostet (angenommen) € 20. Die Kosten eines Falls von klinischer Mastitis, der unter Einsatz von Antiinfektiva behandelt wird, betragen einschließlich des Milchverlustes (angenommen) € 250. Das Präparat (Imrestor®) soll von der Haustierärztin/vom Haustierarzt nur nach Berücksichtigung der Betriebsgegebenheiten verordnet werden. Ab welcher zu erwartenden Inzidenz von klinischer Mastitis in dem genannten Intervall lohnt sich der Einsatz des Präparats?

Angesichts der genannten Vorteile der Entscheidungsanalyse drängt sich die Frage auf, warum sie so gut wie nie angewendet wird.
Folgende möglichen Gründe sind mir (WK) eingefallen:

- In den häufigsten Praxisfällen besteht Zuversicht, die richtige Entscheidung zu kennen
- Abneigung gegen Zahlen (aber: nur die Grundrechenarten werden gebraucht)
- Notwendigkeit, in der Praxis ad hoc zu entscheiden (aber: es ist möglich, eine Entscheidung auch retrospektiv beurteilen)
- Mangel an den notwenigen Informationen
- Entscheidungsanalyse wird als „seelenlos“ empfunden

4. Prophylaxe

4.1. Kausalität in der Medizin

Wenn man Krankheiten verhindern will, muss man offensichtlich versuchen, auf ihre **Ursache** einzuwirken (vgl. Abb. 4.1.-1).

Was aber ist „die Ursache“ einer Krankheit? Der Begriff der Ursache (lateinisch „causa“) und Ursache-Wirkungs-Beziehungen (Kausalität) sind schon seit Jahrtausenden Gegenstand der Philosophie. Aristoteles unterschied vier klassische Ursachen:
causae internae
    causa materialis - das, woraus ein Ding entsteht
    causa formalis - das, woraus ein Ding seine Eigenschaften erhält

causae externae
    causa efficiens - das, was durch sein (äußeres) Wirken das Ding hervorbringt
    causa finalis – das, um dessentwillen ein Ding hervorgebracht wird.

(Warum er auf Latein dachte, bleibt sein Geheimnis.)

Platon fügte noch eine weitere Art von Ursache hinzu: causa exemplaris – Muster, nach dem ein Ding durch eine (vernünftige) c. efficiens hervorgebracht wird.

Wenn man im BROCKHAUS nach einer Definition für Ursache sucht, stößt man auf: „Ding, Zustand oder Geschehen, das notwendig als der reale Grund von anderem anerkannt werden muss“ und kommt darüber etwas ins Grübeln, denn „Grund“ und „Ursache“ sind ja praktisch synonym, so dass das zu Definierende in der Definition auftaucht. Nett.

Kurzes, fast schmerzloses Nachdenken ergibt, dass es verschiedene Sichtweisen der Kausalität gibt. Zunächst kann man unterscheiden, ob man Kausalität als eine von wahrnehmenden Subjekten unabhängige Erscheinung der Natur betrachtet (so genannte ontologisch-realistische Sicht), oder ob man Kausalität als Konstruktion zur Beschreibung und/oder Interpretation von beobachteten Ereignisfolgen und Phänomenen ansieht (methodologisch-nominalistische Sicht). Als nicht begründbare Kategorie, welche Erfahrung erst ermöglicht, wurde die Kausalität von KANT angesehen. Unsere Version lautet: Kausalität ist ein bei vielen Tierarten genetisch verankerter Mechanismus zur Interpretation wahrgenommener Folgen von Ereignissen, der sich in der Evolution hinreichend bewährt hat, auch wenn es zu falschen kausalen Verknüpfungen (Aberglaube) kommen kann.

Ohne philosophische Befrachtung kann man Kausalität auch ganz pragmatisch angehen und mit SUPPES postulieren:
$p(W/U) > p(W)$
Das bedeutet: die Wahrscheinlichkeit, dass ein bestimmtes Ereignis W eintritt, wenn zuvor ein als Ursache angesehenes Ereignis U eingetreten ist, ist größer, als die absolute Wahrscheinlichkeit von W. Anders ausgedrückt: Wenn U (eine) Ursache von W ist, muss die Wahrscheinlichkeit von W unter der Bedingung, dass U eingetreten ist oder vorliegt, größer sein, als die absolute Wahrscheinlichkeit von W. Das sagt natürlich nichts darüber aus, auf welche Weise das Ereignis „U“ das Ereignis „W“ beeinflusst.

Man kann Ursachen (von Krankheiten) unter dem Aspekt betrachten, ob sie notwendig und/oder ausreichend sind.

Tab. 4.1-1: Einteilung von Ursachen ($\Rightarrow$ bedeutet „daraus folgt“, $\neg$ bedeutet „kein“)

	notwendig	nicht notwendig
ausreichend	$P(W/U) = 1$ $U \Rightarrow W$	$U \Rightarrow W$ $X_{1(-n)} \Rightarrow W$
nicht ausreichend	$\neg U \Rightarrow \neg W$ U und $X_{1(-n)} \Rightarrow W$	U und $X_{1(-n)} \Rightarrow W$ $X_{1(-n)} \Rightarrow W$

Zu den notwendigen und ausreichenden Ursachen von Krankheiten zählen die Erreger klassischer Infektionskrankheiten. Sie waren der Ausgangspunkt für die HENLE-KOCHschen Postulate:

1. Der „Parasit“ ist in jedem Fall der Krankheit anzutreffen und zwar unter Verhältnissen, welche den pathologischen Veränderungen und dem klinischen Verlauf der Krankheit entsprechen.
2. Er kommt bei keiner anderen Krankheit als zufälliger oder nicht pathogener Schmarotzer vor.
3. Er kann vom Körper isoliert in Reinkulturen hinreichend oft umgezüchtet werden und ist danach imstande, von neuem die Krankheit auszulösen.

In die Kategorie „notwendig, aber nicht ausreichend“ fallen etwa Mikroorganismen bei den so genannten infektiösen Faktorenerkrankungen. Ohne Anwesenheit dieser „Erreger“ kommt die Krankheit nicht zustande, aber ihre Anwesenheit allein genügt nicht zur Auslösung.

Kann eine Krankheit durch verschiedene Faktoren jeweils sicher ausgelöst werden, so ist jeder einzelne dieser Faktoren als nicht notwendige, aber ausreichende Ursache anzusehen.

Die letzte Kategorie wird von „Ursachen“ gebildet, welche weder allein in der Lage sind, eine Krankheit auszulösen, noch notwendig sind, weil die betreffende Krankheit auch von ganz anderen Faktoren hervorgerufen werden kann.

EVANS hat 1976 die HENLE-KOCHschen Postulate unter modernen epidemiologischen Gesichtspunkten erweitert:

1. Die Prävalenz der Erkrankung sollte bei Individuen, welche gegenüber der vermuteten Ursache exponiert waren, signifikant höher sein als bei denen, die nicht exponiert waren.
2. Exposition gegenüber der vermuteten Ursache sollte sich bei Individuen mit der betreffenden Krankheit signifikant häufiger nachweisen lassen als bei Individuen ohne die Krankheit, wenn alle anderen Risikofaktoren konstant sind.
3. In prospektiven Studien sollte sich die Inzidenz der Krankheit bei exponierten Individuen als signifikant höher erweisen als bei nicht exponierten.
4. Die Krankheit sollte nach der Exposition in Erscheinung treten, wobei die Inkubationszeit eine Normalverteilung zeigen sollte.
5. Die Reaktion bei den exponierten Individuen sollte ein Spektrum von mild bis schwer umfassen.
6. Bei den Exponierten sollte mit großer Wahrscheinlichkeit eine messbare Reaktion auftreten, welche vorher nicht nachweisbar war (z.B. Antikörper oder Krebszellen), oder

sich verstärken, wenn sie vorher vorhanden war. Dieses Reaktionsmuster sollte bei nicht exponierten Individuen selten sein.

7. Die experimentelle Auslösung der Krankheit sollte bei exponierten Tieren oder Menschen häufiger gelingen als bei nicht exponierten. Diese Exposition kann bei Freiwilligen absichtlich herbeigeführt werden, im Laborexperiment erfolgen, oder durch Beeinflussung natürlicher Exposition erfolgen.
8. Elimination oder Abschwächung der Exposition gegenüber einer vermuteten Ursache sollte die Inzidenz senken.
9. Verhinderung oder Abschwächung der Reaktion auf die vermutete Ursache sollte die Krankheit abschwächen oder verhindern.
10. Alle Beziehungen und Befunde sollten biologisch und epidemiologisch plausibel sein.

4.2 Maße für Häufigkeit von Krankheiten

Wenn prophylaktische Maßnahmen wirksam sind, müssen sie die Häufigkeit der in Frage stehenden Krankheit senken. Auf welche Weise soll aber die Häufigkeit einer Krankheit gemessen werden? In diesem Abschnitt werden die beiden Begriffe Prävalenz und Inzidenz erklärt.

Der Begriff der **Prävalenz** tauchte schon im Kapitel 2.3 als (momentaner) Anteil der Individuen mit einer bestimmten Krankheit an einer bestimmten Population. In der Vierfeldertafel (s. Tab. 2.3.-2) entsprach sie dem Bruch (a + c)/(a + b + c + d).

In Abbildung 4.2-1 sind schematisch Ereignisse bei 10 Individuen (A bis J) über einen gewissen Zeitraum (12 gleiche Zeitabschnitte, zum Beispiel Monate) dargestellt. Die roten Bereiche sollen Episoden einer bestimmten Krankheit darstellen. Individuum B ist schon zu Beginn des Zeitraums, der untersucht werden soll (t0), erkrankt. Insgesamt 7 Individuen erkranken im fraglichen Zeitraum neu. Zum Zeitpunkt t1 ist kein Individuum krank. Hier ist die Prävalenz 0/10 = 0. Dagegen sind zum Zeitpunkt t4 4 Individuen krank. Die Prävalenz beträgt daher 4/10 = 0,4.

Die **Inzidenz** gibt an welcher Anteil der Individuen unter Risiko in einem bestimmten Zeitraum neu erkrankt. Unter Risiko stehen die Individuen, welche die Krankheit bekommen können (so können beispielsweise nur weibliche Tiere Endometritis bekommen) und noch nicht gehabt haben (grüne Felder). Im gewählten Beispiel sind das 9 Individuen (Individuum B scheidet aus, weil es schon zu Beginn des betrachteten Zeitraums erkrankt war.), von denen 7 erkranken. Das entspricht einer **kumulativen Inzidenz** von 7/9 = 0,78. Prävalenz und kumulative Inzidenz können demnach nur Werte zwischen 0 und 1 annehmen, wenn sie als Dezimalbruch angegeben werden.

Bei Krankheiten mit kurzem Verlauf (und dem Ergebnis Heilung oder Tod) kann die Prävalenz zu vielen Zeitpunkten deutlich geringer als die kumulative Inzidenz sein. Umgekehrt kann die Prävalenz von Krankheiten mit chronischem Verlauf deutlich höher als die Inzidenz sein.

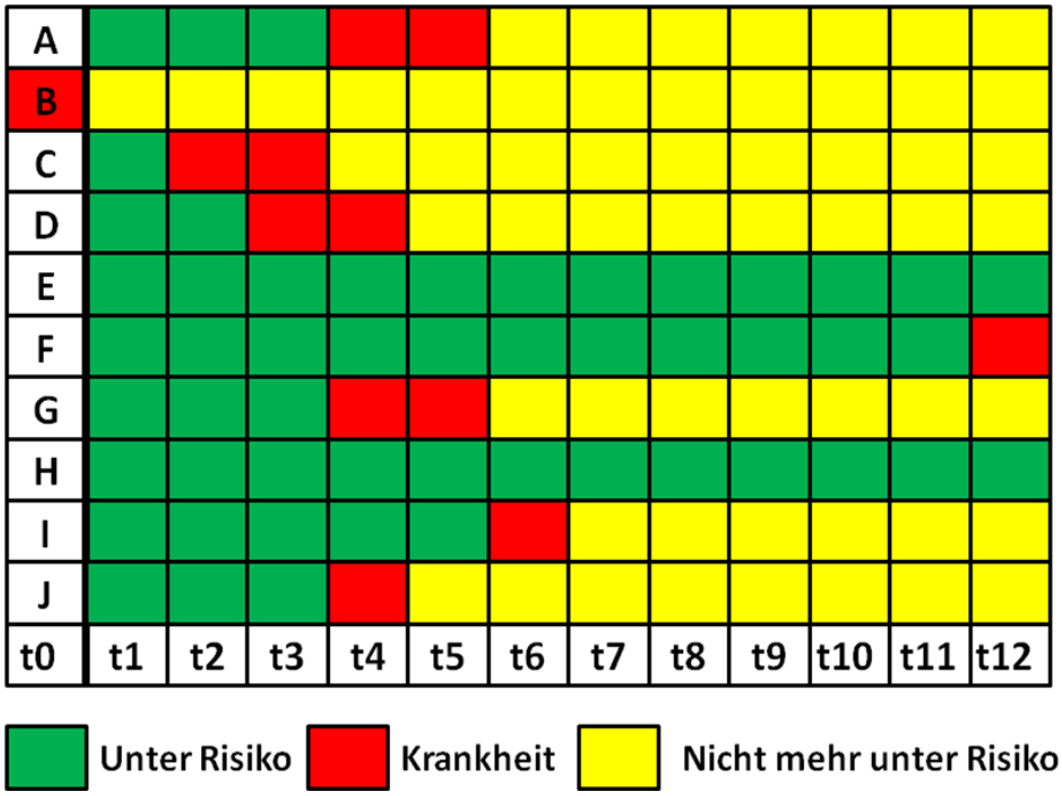


Abb. 4.2-1: Schematische Darstellung zu Prävalenz und Inzidenz

Bei der Berechnung der **Inzidenzdichte** wird berücksichtigt, wie lange die Individuen innerhalb des betrachteten Zeitraums unter Risiko standen. Unter Risiko stehen 52 „Individuum-Monate“ (grüne Felder) oder 52/12 = 4,33 Individuum-Jahre. In dieser Zeit sind 7 Fälle neu aufgetreten. Die Inzidenzdichte beträgt dann 7/4,33 = 1,62. Pro Individuum-Jahr unter Risiko sind also etwa 1,6 Fälle der Erkrankung neu aufgetreten. Die Inzidenzdichte kann Werte zwischen 0 und ∞ annehmen.

4.3 Wahrscheinlichkeit

Wir verwenden wohl täglich den Begriff der Wahrscheinlichkeit. Aber was ist das denn, Wahrscheinlichkeit? Gibt es verschiedene Arten von Wahrscheinlichkeiten? (Erst überlegen, dann weiterlesen!)

Man kann folgende Arten von Wahrscheinlichkeiten unterscheiden:

- logische Wahrscheinlichkeit
- statistische, empirische oder frequentistische Wahrscheinlichkeit
- subjektive Wahrscheinlichkeit

Zunächst zur logischen Wahrscheinlichkeit. Wenn man eine Münze wirft, ist es relativ leicht einzusehen, dass jedem der beiden möglichen Ergebnisse die gleiche Wahrscheinlichkeit von 0,5 zukommt. Ähnliches gilt für einen Würfel, für ein Roulette und für die Kugeln in der Lottotrommel, aber auch etwa für einfache Erbgänge. Wenn die zu Grunde liegenden Mechanismen bekannt sind, können die Wahrscheinlichkeiten rein rechnerisch ermittelt werden.

Die zweite Art von Wahrscheinlichkeit kann statistisch, empirisch oder frequentistisch genannt werden. Wenn eine retrospektive Analyse ergibt, dass 15 % aller Patienten mit Krankheit A das Symptom S1 gezeigt haben, wird die Wahrscheinlichkeit, dass ein weiterer Patient mit Krankheit A das Symptom S1 zeigt, mit 0,15 oder 15 % angesetzt. Diese Folgerung ist streng genommen nicht ganz korrekt, aber üblich. Sie wäre korrekt, wenn der in Frage stehende Patient der Grundgesamtheit angehörte, aus der die diagnostische Sensitivität von Symptom S1 ermittelt wurde.

Diese Art von Wahrscheinlichkeit unterscheidet sich ganz eindeutig von der zuerst genannten. Keine noch so umfangreiche Kenntnis biomedizinischer Mechanismen erlaubt eine Aussage über die Häufigkeit eines Symptoms.

Eine dritte Art von Wahrscheinlichkeit ist die subjektive. Wie groß schätzen Sie die Wahrscheinlichkeit, dass ein bestimmter Tennisspieler im nächsten Turnier Wimbledon gewinnt? Logische oder mathematische Überlegungen helfen uns hier nichts, und die Erfahrung nicht viel mehr. Denn es ist offensichtlich unsinnig, von dem Anteil seiner Siege an den insgesamt bisher durchgeführten Turnieren auszugehen. Von den Unterschieden in der subjektiven Wahrscheinlichkeit verschiedener Menschen leben die Buchmacher. Und von dem Unvermögen der Menschen, mit sehr geringen Wahrscheinlichkeiten umzugehen, lebt unter anderem das LOTTO 6 aus 49 (das von einem Präsidenten der Deutschen Bundesbank einmal als „Sondersteuer für Idioten“ bezeichnet worden sein soll). Zum Lotto: Die Wahrscheinlichkeit, den Jackpot mit einem Spielfeld zu knacken, also 6 Richtige und richtige Superzahl zu haben, beträgt 1:139.838.160. (Herleitung: Die Wahrscheinlichkeit des gemeinsamen Auftretens voneinander unabhängiger Ereignis ist gleich dem Produkt der Einzelwahrscheinlichkeiten. Die Wahrscheinlichkeit, die erste Zahl richtig zu haben, beträgt 6/49. Die Wahrscheinlichkeit, auch die zweite Zahl richtig zu haben, beträgt 5/48 usw. bis 1/44 bei der sechsten Zahl. Die Chance, die Superzahl korrekt zu haben, beträgt offensichtlich 1/10. Auch wenn manche Menschen glauben, dass Lottozahlen einem System gehorchen – und sich arm gespielt haben – ist das nicht so, und die Trommel hat kein Gedächtnis. Die Kombination 1,2,3,4,5,6 hat die gleiche Chance wie jede 3andere.) Zur Illustration stelle man

sich eine Strecke von 1400 km vor und suche den richtigen Zentimeter heraus. Trotzdem wird der Jackpot immer wieder geknackt. Auch äußerst unwahrscheinliche Ereignisse treten ein. Dabei ist aber zu berücksichtigen, dass Millionen von Spielern teilnehmen. Wenn 50 Millionen Spielfelder in voneinander unabhängiger Weise ausgefüllt werden, ist die Wahrscheinlichkeit, dass der Jackpot mindestens einmal geknackt wird, etwa 0,3 oder 30 %.

Ist Wahrscheinlichkeit etwas, was es gibt, oder existiert sie nur in unserer Vorstellung? In einem von Menschen unberührten Teich überleben längst nicht alle Kaulquappen, aber die Wahrscheinlichkeit, dass genügend zu geschlechtsreifen Fröschen werden, welche die Population erhalten, ist **hinreichend** groß. Es ist nicht anzunehmen, dass die Frösche oder ihre Fressfeinde Wahrscheinlichkeitsberechnungen anstellen Ein anderes Beispiel: Die Raupe des Kaiseratlas, eines Schmetterlings, verpuppt sich in einem Blatt eines bestimmten Strauches, dessen Stiel sie annagt, wonach das Blatt sich einrollt und braun wird. Die Raupe nagt aber die Stiele mehrerer Blätter an, bevor sie sich tatsächlich in einem verpuppt. Würde sie nur den Stiel eines Blattes annagen, würde das eine braune Blatt unter all den grünen Blättern sicher die Aufmerksamkeit von Vögeln erregen, die nach der Belohnung durch den darin enthaltenen Leckerbissen rasch hinter das Geheimnis der braunen Blätter kommen würden. Die Wirksamkeit der Strategie der Raupe setzt voraus, dass sie so viele Blattstiele annagt, dass Vögel mit für den Zweck des Überlebens der Puppe hinreichender Wahrscheinlichkeit mehrere Fehlversuche erleben und so von weiteren Versuchen ablassen.

Für einen bestimmten Zweck hinreichende Wahrscheinlichkeit ist ein äußerst wichtiger Aspekt.

4.4 Studien zur Quantifizierung des Zusammenhangs zwischen Faktoren und Krankheiten

Vor Beginn einer Studie sollte so genau wie möglich festgelegt werden, welche Frage mit „ja/nein“ oder mit einer Zahl beantwortet werden soll.

Wenn geprüft werden soll, ob Faktor F ursächliche Bedeutung bei der Mysteriose der Rinder zukommt, gibt es verschiedene Wege, zu einer Antwort zu kommen. Die einfachste Möglichkeit besteht darin, eine so genannte ***Querschnittsuntersuchung*** durchzuführen. Dabei wird eine größere Population im Hinblick auf das Vorliegen der Krankheit und des Faktors F untersucht. Die damit verbundenen theoretischen und praktischen Schwierigkeiten sollen wir zunächst außer Acht gelassen werden. Die Ergebnisse lassen sich – wie könnte es anders sein – in einer Vierfeldertafel zusammenfassen (s. Tab. 4.4-1)

Tab. 4.4-1: Vierfeldertafel zur Darstellung der Verteilung von Individuen mit und ohne Mysteriose und mit und ohne Faktor F

	Mysteriose	**Keine Mysteriose**
Faktor F vorhanden	**A**	**B**
Faktor F nicht vorhanden	**C**	**D**

Angenommen, es werden 1000 Kühe untersucht und dabei 80 Fälle von Mysteriose entdeckt. Faktor F war bei 400 Kühen vorhanden (s. Tab. 4.4-2).

Tab. 4.4-2: Vierfeldertafel zur Darstellung der empirisch ermittelten Verteilung von 1000 Individuen mit und ohne Mysteriose und mit und Faktor F

	Mysteriose	**Keine Mysteriose**	
Faktor F vorhanden	**A**	**B**	**400**
Faktor F nicht vorhanden	**C**	**D**	**600**
	80	**920**	**1000**

Querschnittsuntersuchungen liefern keinen Hinweis auf die Ereignisfolge, sondern lediglich auf eine Assoziation.

Wenn kein Zusammenhang zwischen F und Mysteriose besteht, die beiden Kategorien also unabhängig voneinander sind. wäre zu erwarten, dass in Feld A 1000*0,08 * 0,4 = 32 Kühe mit Mysteriose und Faktor F zu finden sind. Die Inhalte der übrigen Felder können analog errechnet werden. Die Ergebnisse sind in Tab. 4.4-3 wiedergegeben.

Tab. 4.4-3: Aufgrund des Zufalls zu erwartende Verteilung von 1000 Individuen mit und ohne Mysteriose und mit und ohne Faktor F

	Mysteriose	**Keine Mysteriose**	
Faktor F vorhanden	**32**	**368**	**400**
Faktor F nicht vorhanden	**48**	**552**	**600**
	80	**920**	**1000**

Die tatsächlich gefundene Verteilung ist in Tab. 4.4-4 dargestellt.

Tab. 4.4-4: Tatsächliche Verteilung von 1000 Individuen mit und ohne Mysteriose und mit und ohne Faktor F

	Mysteriose	**Keine Mysteriose**	
Faktor F vorhanden	**62**	**338**	**400**
Faktor F nicht vorhanden	**18**	**582**	**600**
	80	**920**	**1000**

Eine Maßzahl zur Berechnung der Stärke der Assoziation zwischen Faktor F und Mysteriose ist die so genannte ***Odds Ratio*** (Chancenverhältnis). In der nachstehenden Formel (17) bezieht sich A, B, C und D auf die Vierfeldertafel in Tab. 4.4-2.

$$\textbf{Odds ratio} = A/C/B/D = AD/BC \qquad (17)$$

Das Verhältnis der beiden Quotienten ist 1,0, wenn die Verteilung auf die Felder den Erwartungen entspricht (s. Tab. 4.4-3). Es ist größer als 1, wenn der Faktor das Entstehen der Krankheit begünstigt („Risikofaktor“) und kleiner als 1, wenn der Faktor die Wahrscheinlichkeit für die Entstehung der Krankheit vermindert. In dem in Tab. 4.4-4 dargestellten Fall beträgt das Chancenverhältnis 5,9. Das bedeutet, dass die Verteilung von der auf Grund des Zufalls zu erwartenden deutlich abweicht.
Zur Prüfung der Frage, ob die Abweichung im statistischen Sinn signifikant ist, berechnet man das 95 % oder 99 % Vertrauensintervall für die Odds Ratio. Schließt es die Zahl 1 ein, kann nicht mit der gewünschten Sicherheit entschieden werden, ob der verdächtigte Faktor einen Einfluss auf das Entstehen der fraglichen Krankheit hat.
Wenn für alle „verdächtigen“ Faktoren z.B. (Futterkomponenten, Tränke, Impfungen oder andere Behandlungen, Stallabteilung, Altersgruppe etc.) diese Berechnungen angestellt werden, sollten sich gewisse Kandidaten zu erkennen geben. Wenn das nicht in recht eindeutiger Weise der Fall ist, liegt der Verdacht nahe, dass es noch andere relevante Faktoren gibt, oder dass eine Kombination von Faktoren „am Werk“ war/ist.
Mit der OR lässt sich aber auch ermitteln, wie sich das Risiko, exponiert gewesen zu sein zwischen Erkrankten und nicht Erkrankten verhält: OR = (A:C)/(B:D).

Anwendung des Chi-Quadrat-Testsergibt, dass die Verteilung auf die Felder in hochsignifikanter Weise von der theoretisch zu erwartenden abweicht. Es gibt noch andere Betrachtungsweisen. Unter den Individuen mit Faktor F beträgt die Prävalenz von Mysteriose 62/400 = 0,155; unter denen ohne Faktor F beträgt sie 18/600 = 0,03. Da es sich ja um jeweils

nur eine Stichprobe handelt, liegt die Frage nahe, wie groß die wahren Prävalenzen sind. Das lässt sich natürlich nicht genau sagen, aber es ist möglich, ein Intervall anzugeben, innerhalb dessen diese Werte mit einer vorgegebenen Wahrscheinlichkeit (üblich sind entweder 95 oder 99 %) liegen. Mit Hilfe entsprechender Programme lassen sich diese so genannten Konfidenzintervallee (Vertrauensbereiche) leicht ermitteln. Das 95 % Konfidenzintervall liegt im ersten Fall zwischen 0,121 und 0,194 und im zweiten Fall zwischen 0,018 und 0,047. Die beiden Intervalle überlappen sich also nicht. Auch das ist ein Hinweis, dass die Verteilung der Individuen auf die vier Felder nicht zufällig ist.

Ist nun die Ursache und damit den Schlüssel zur Verhinderung von Mysteriose, dieser Geißel der Rinderhaltung, gefunden? Leider gibt es da ein paar hässliche kleine Probleme. Wie zuverlässig ist die Zuteilung der Individuen? War denn die Diagnose „Mysteriose“ in jedem Fall eindeutig? Sind 1000 unsichere Diagnosen besser als 10? Und was war zuerst da, der Faktor oder die Krankheit? Auch eine statistisch gesicherte Assoziation beweist noch keinen kausalen Zusammenhang und gibt erst recht keinen Hinweis auf die Richtung eines solchen Zusammenhangs. Ratten sind nicht die Erreger von Mülldeponien.

Allerdings ist ein kausaler Zusammenhang recht unwahrscheinlich, wenn sich bei Untersuchung einer großen Stichprobe keine statistisch gesicherte Assoziation ergibt. (Von Spitzfindigkeiten wollen wir jetzt einmal absehen.)

Unter den 400 Kühen mit Faktor F ist die Prävalenz von Mysteriose mit 0,155 zwar mehr als 5mal so hoch wie bei den übrigen (0,03), aber der allergrößte Teil der Kühe mit Faktor F sind (zumindest zum Zeitpunkt unserer Untersuchung) frei von Mysteriose. Das spricht sehr dafür, dass der Faktor F so gefährlich im Hinblick auf die Entstehung von Mysteriose nicht sein kann.

Die relative Bedeutung eines Faktors (im Vergleich zu anderen) für das Auftreten einer bestimmten Krankheit kann mit der so genannten logistischen Regression berechnet werden.

Ein besseres Verfahren als eine Querschnittsuntersuchung stellt die so genannte ***Fall-Kontroll-Studie*** dar. Ihr Prinzip ist in Abbildung 4.4-1 dargestellt. Es werden Fälle der fraglichen Krankheit und eine angemessene Zahl von Kontrollindividuen daraufhin untersucht, inwieweit in der Vergangenheit Exposition gegenüber dem fraglichen Faktor stattgefunden hat. Wenn der Faktor zur Entstehung der Krankheit beiträgt, sollte er bei den Individuen mit der Krankheit deutlich öfter im gezielt erhobenen Vorbericht auftauchen als bei den Kontrollen. Im dargestellten Beispiel wurde die Einwirkung des Faktors bei ¾ der Fälle, aber nur bei ¼ der Kontrollen festgestellt.

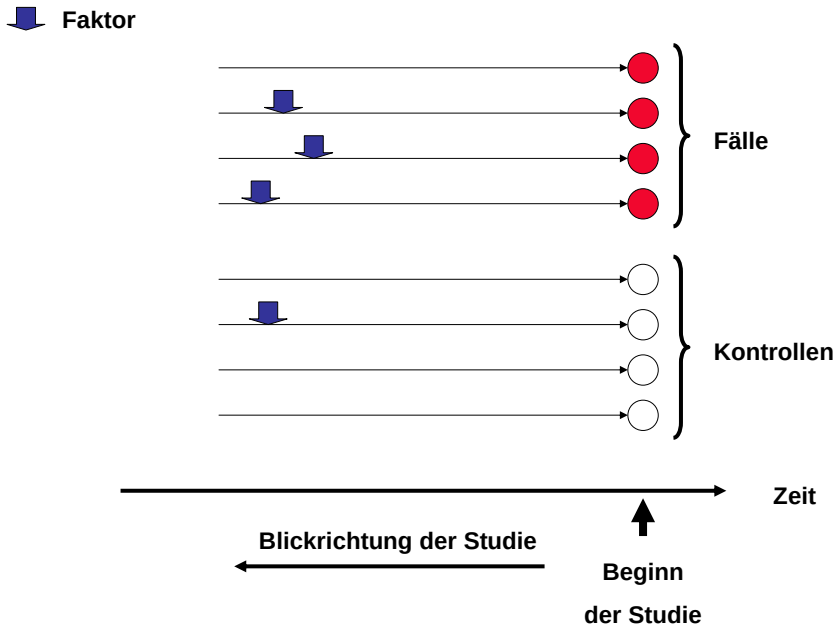
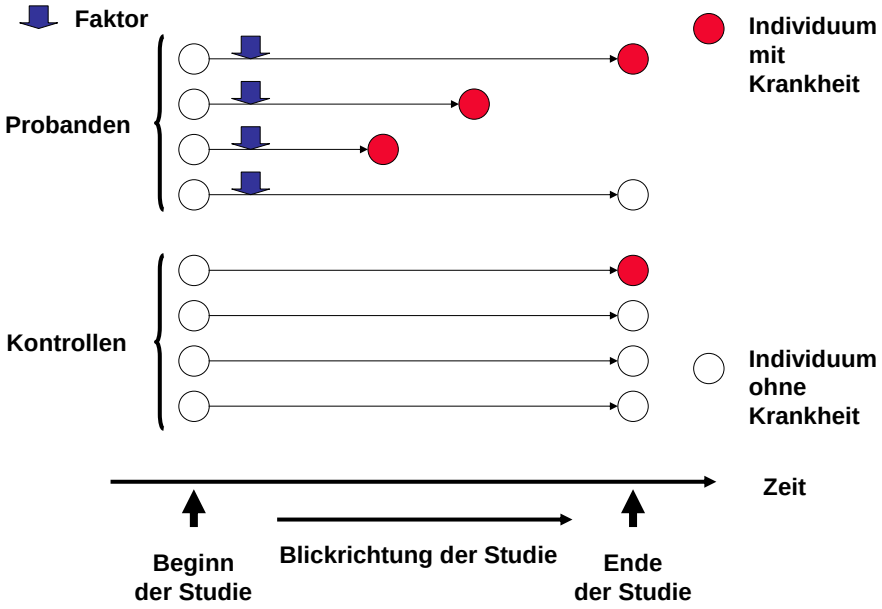


Abb.: 4.4-1: Schematische Darstellung des Prinzips einer Fall-Kontroll-Studie

Wie fast immer in der wirklichen Welt, sind die Verhältnisse nicht so ganz klar, wie man sich das wünschen würde, denn es gibt da einen Teil der Fälle, bei denen die Einwirkung des Faktors nicht ermittelt werden konnte. Hierfür gibt es mehrere mögliche Erklärungen. So kann die Kategorisierung als Fall falsch gewesen sein, die Einwirkung des Faktors hat tatsächlich stattgefunden, wurde aber nicht ermittelt, oder es gibt noch andere ausreichende Ursachen zur Auslösung der Krankheit. Und wie steht es mit den Kontrollen, bei denen eine Einwirkung des Faktors in dem zurückliegenden Zeitraum, über den sich die retrospektive Studie erstreckt, stattgefunden hat? Neben der offensichtlichen Erklärungsmöglichkeit, dass ein Irrtum hinsichtlich der Einwirkung des Faktors vorliegt, könnte es je nach Art der Krankheit auch zu einer Heilung gekommen sein. Oder die Intensität der Einwirkung des Faktors hat nicht ausgereicht, die Krankheit auszulösen.

Es zeigen sich hier Probleme, die im Bereich der Epidemiologie immer wieder auftauchen: zum einen müssen von der sehr komplexen Realität vereinfachte Abstraktionen erstellt werden, damit eine Bearbeitung stattfinden kann, was aber mit Verlust an Information erkauft wird. Zum anderen verbleibt immer eine Unsicherheit, die zwar quantifizierbar, aber nicht auszuschalten ist.

Die direkteste Methode zur Ermittlung der Bedeutung eines Faktors bei der Entstehung einer Krankheit ist die kontrollierte prospektive ***Kohortenstudie***. Ihr Prinzip ist in Abb. 4.4-2 schematisch dargestellt. Sie geht aus von vergleichbaren Gruppen von Probanden und Kontrollen. Nur die Probanden werden der Einwirkung des zu prüfenden Faktors ausgesetzt und es wird kontrolliert, welche Anteile der Gruppen in einem bestimmten Zeitraum (dessen Länge von der Art der Krankheit abhängt) die Krankheit bekommen.



Ab. 4.4-2: Schematische Darstellung des Prinzips einer kontrollierten prospektiven Kohortenstudie

Als Maß für den Zusammenhang zwischen Exposition gegenüber dem Faktor und dem Auftreten der Krankheit wird hier das so genannte relative Risiko berechnet. In der nachstehenden Formel (18) beziehen sich A- D auf die Vierfeldertafel in Tab. 4.4-1.

Relatives Risiko = A/(A+B):C/(C+D) **(18)**

Das relative Risiko entspricht also dem Verhältnis der Inzidenzen der beiden Gruppen (Probanden bzw. Kontrollen). In dem Beispiel aus Abb. 4.4-2 würde sich dabei 3 ergeben (3/4 geteilt durch ¼).

Nicht immer ist es praktisch möglich und ethisch vertretbar, eine prospektive Kohortenstudie durchzuführen. In manchen Fällen ist es möglich, retrospektiv eine solche Studie anhand von Archivdaten durchzuführen. Dann liegt der Beginn des Untersuchungszeitraums bei den einzelnen Patienten und Kontrollen mehr oder weniger weit in der Vergangenheit, der Endpunkt jedoch stets vor Beginn der Studie.

Bei Studien zum Einfluss eines Faktors auf das Auftreten einer Krankheit ist auf so genannte „Confounder“ zu achten. Das sind Gegebenheiten (z.B. Kolostrumversorgung, Alter, Aufstallung, Impfungen), welche indirekt die Zielgröße (also Auftreten einer Krankheit beeinflussen und so den Einfluss des untersuchten Faktors (Anwesenheit eines bestimmten Erregers) vernebeln. Ein häufig zitiertes Beispiel ist die Tatsache, dass die Zahl der Storchennester und die mittlere Kinderzahl in einer Region korrelieren. Der „Confounder“ ist hier die „Ländlichkeit“ der Region.

Die Nichtexistenz eines Einflusses oder die Unmöglichkeit eines Ereignisses kann nicht bewiesen werden. Beispiel: In einer Studie über den möglichen Einfluss von Mobilfunkfeldern auf die Gesundheit und Leistung von Rindern wäre es aus Sicht der

Öffentlichkeit wünschenswert, dass die absolute Unschädlichkeit dieser Felder bewiesen würde. Das ist aber nicht möglich.

5. Subjektive Aspekte

5.1 Wahrnehmung

Wahrnehmung ist bei der klinischen Untersuchung offensichtlich von großer Bedeutung. Sie wird von einer Reihe von Faktoren beeinflusst, welche von dem wahrgenommenen Gegenstand oder Ereignis unabhängig sind und zum Teil im wahrnehmenden Subjekt begründet sind. Dazu gehören unter anderem Sinnesschärfe, Training, Aufmerksamkeit/Ablenkung, Licht- und Lärmverhältnisse, Gruppendruck und Erwartung (s. Abb. 5.1-1).

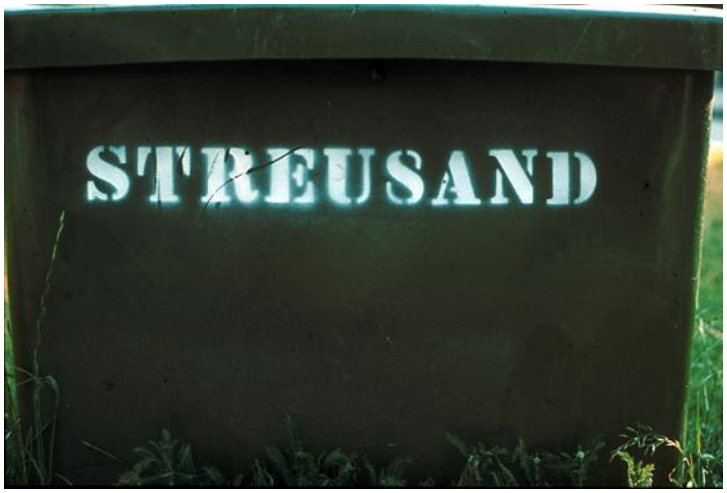


Abb. 5.1-1 Beispiel für den Einfluss der Erwartung auf die Wahrnehmung (Für Leser, die mit dieser Abbildung nichts anfangen können: Das Zeichen zwischen „U“ und „A“ wird als „S“ interpretiert, weil dieser Buchstabe dort erwartet wird, ist aber die Ziffer „8“.)

5.2 Informationsverarbeitung

Das menschliche Gehirn (ein Klumpen organischer Materie, die über sich selbst nachdenken kann!) verarbeitet ständig sehr viele Informationen. Zum größten Teil bleibt das aber dem Bewusstsein verborgen. Bewusst können die meisten Menschen aber nur eine sehr begrenzte Anzahl von Informationseinheiten (etwa 7) gleichzeitig verarbeiten. Werden sie mit mehr Einheiten überladen, sinkt die Qualität des Ergebnisses.

5.3 (Klinische) Erfahrung

Der Begriff „Erfahrung“ wird von Klinikern häufig verwendet („Wie die klinische Erfahrung zeigt ...“), ohne dass sie (sich) im Einzelfall Rechenschaft darüber abgeben, was genau damit gemeint ist.

Die in Enzyklopädien angebotenen Definitionen (z.B.: „Inbegriff von Erlebnissen in einem geordneten Zusammenhang, ebenso die in ihnen gegebenen Gegenstände und die durch sie erworbenen Kenntnisse und Fähigkeiten“ BROCKHAUS-Enzyklopädie, 19. Auflage) helfen nicht viel weiter.

Es scheint Menschen angeboren zu sein, alle Beobachtungen, welche mit ihren Vorstellungen in Einklang zu bringen sind, als Beweise für die Richtigkeit ihrer Vorstellungen zu werten. Zwei in einem Gebiet tätige Tierärzte können ganz unterschiedliche Überzeugungen haben, die sie aber beide mit ihren Erfahrungen begründen. Selbst bei Tierärzten in Kliniken, die jahrelang dieselben Patienten gesehen haben, können unterschiedliche Überzeugungen bestehen. Dieses Phänomen ist jedoch nicht spezifisch tierärztlich. Auf jeden Fall aber ist Skepsis gegenüber Behauptungen angebracht, welche allein mit Erfahrung begründet werden.

Andererseits ist natürlich ohne Erfahrung keine sinnvolle Tätigkeit möglich. Sie ist unverzichtbar.

Klinische Erfahrung kann sich unter anderem auf folgende Größen beziehen:
- lokale Inzidenzen und Prävalenzen von Krankheiten (dies dürfte ein sehr wichtiger Teil der klinischen Erfahrung sein!),
- diagnostische Sensitivität von Symptomen bei bestimmten Krankheiten,
- Wahrscheinlichkeit von Ereignissen nach bestimmten Interventionen bei bestimmten Krankheiten (konkreter: Erfolgsaussichten bestimmter therapeutischer Interventionen).

5.4 Fehldiagnosen

Wozu ein Kapitel über Fehldiagnosen? Fehldiagnosen stellen nur die anderen. Oder?
Es ist nur scheinbar einfach zu definieren, was unter einer Fehldiagnose zu verstehen ist, denn das hängt unter anderem von den Ansprüchen ab, die man an eine Diagnose stellt. Gewiss gibt es klare Fälle. Wenn ein Tierarzt bei einer Kuh die Diagnose „Tollwut“ stellt, die Untersuchung des Gehirns aber „Hirnstammenzephalitis vom Typ der Listeriose“ und keinen Hinweis auf Tollwut ergibt, hat er offensichtlich eine Fehldiagnose gestellt. Wir schlagen vor, den Begriff der „klinisch relevanten Fehldiagnose“ einzuführen und definieren ihn als nicht zutreffende Zuordnung eines konkreten Erkrankungsfalles zu einer Kategorie eines nosologischen Systems, wenn die eigentliche Krankheit eine andere Intervention erfordert hätte, und der Patient und/oder sein Besitzer dadurch zusätzlichen gesundheitlichen bzw. finanziellen Schaden erlitten hat. Dies ist im genannten Beispiel offensichtlich der Fall.

Mängel in der Untersuchung sind mit Sicherheit die wichtigste Ursache für Fehldiagnosen, wie der folgende Spruch zum Ausdruck bringt: More is missed by not looking than by not knowing. Vorschnelle Festlegung auf eine Diagnose kann dazu führen, dass die Untersuchung

nur noch in diese Richtung fortgeführt wird, wobei wichtige Befunde, die in eine andere Richtung weisen, nicht erhoben oder ignoriert werden. Die Bedeutung der klinischen Untersuchung kann gar nicht oft genug betont werden. Ein zweiter wichtiger Aspekt ist der Umfang des diagnostischen Repertoires des Untersuchers. Wer bei jeder Kuh, die nicht frisst, eine „Indigestion“ oder bei jedem Schwein mit erhöhter Körpertemperatur einen „fieberhaften Infekt“ diagnostiziert, weil er sonst keine Krankheiten kennt, wird natürlich häufig „richtig liegen“. Allgemein ausgedrückt: je breiter eine „Diagnose“ gefasst ist, umso größer ist die Wahrscheinlichkeit, dass sie „zutreffend“ ist. Der Nachteil dieser Methode ist evident, wenn mit dieser Diagnose verschiedene Krankheiten „abgedeckt“ werden, die aus prognostischer, therapeutischer oder seuchenhygienischer Hinsicht unterschiedlich zu beurteilen sind.

Dies wirft die Frage nach der sinnvollen „Tiefe“ von Diagnostik auf. Sie hängt vor allem von den verfügbaren Interventionen ab, wie folgender Spruch demonstriert, der von alten amerikanischen Rinderpraktikern überliefert ist: „Stehende Kühe werden mit Sulfonamiden, liegende Kühe werden mit Kalzium behandelt.“ Unter praktischen Gesichtspunkten ist Diagnostik also so weit sinnvoll, wie noch Handlungsalternativen offenstehen. Dies kann bei verschiedenen Tierarten unterschiedlich sein. So reicht in der Buiatrik die Diagnose „chronisches Herzversagen“ meist aus, eine Entscheidung zu fällen, während in der Kleintiermedizin noch verschiedene Möglichkeiten mit unterschiedlicher Prognose abzuklären sind.

6. Antworten auf die Fragen und Übungsaufgaben im Text

2.3.2-1: Die Annahme ist nur gerechtfertigt, wenn sich die Individuen mit der bestimmten Krankheit hinsichtlich des Tests homogen verhalten. Das ist jedoch nicht immer der Fall. So ist es möglich, dass beim Ausbruch einer Infektionskrankheit sich zunächst fast alle betroffenen Individuen im akuten Stadium befinden, in dem Fieber regelmäßig auftritt, während später nur noch wenige Individuen betroffen sind, von denen sich fast alle im chronischen Stadium befinden, in dem kein Fieber mehr besteht (Beispiel: Salmonellose). Allgemeiner ausgedrückt: wenn es unter den Individuen, welche mit einem bestimmten Krankheitsetikett belegt werden, Gruppen gibt, die sich hinsichtlich des Testausfalls inhomogen verhalten, hat dieser Test keine konstante diagnostische Sensitivität für diese Krankheit (wenn verschiedene Testpopulationen hinsichtlich der Untergruppen unterschiedlich zusammengesetzt sind).

2.3.2-2: Sofern die Stichprobe der Kranken repräsentativ für die Population der Individuen mit K ist, wird die diagnostische Sensitivität richtig ermittelt. Die diagnostische Spezifität wird bei Verwendung von „Gesunden“ aber möglicherweise zu hoch eingeschätzt, weil bei ihnen (falsch) positive Testergebnisse seltener auftreten als bei Individuen mit differentialdiagnostisch zu berücksichtigenden Krankheiten.

2.3.3-1: Hohe diagnostische Sensitivität ist wichtig, wenn möglichst alle Individuen mit einer bestimmten Krankheit aus einer Population „herausgefischt“ werden sollen. Das ist sinnvoll beim „Screening“ für eine Krankheit, die ernst, aber (besonders im Frühstadium) heilbar (oder zumindest behandelbar) ist, zum Beispiel Diabetes mellitus. Eine weitere Situation, in der es auf möglichst hohe diagnostische Sensitivität ankommt, ist der Versuch der Ausrottung einer Krankheit. Beispiel: Tuberkulose und Brucellose bei Rindern nach dem zweiten Weltkrieg. Wenn die Tests nicht gleichzeitig absolut spezifisch sind, was kaum zu erwarten ist, wird unter den Testpositiven eine mehr oder weniger große Anzahl Individuen mit falsch positiven Ergebnissen sein. Im Falle von Diabetes mellitus können sie weiteren, spezifischeren Tests unterzogen werden, im Falle der genannten Rinderkrankheiten wurden diese „Justizmorde“ als Kosten der Bekämpfung in Kauf genommen.

Ein Test mit möglichst hoher diagnostischer Sensitivität ist auch wünschenswert, wenn eine Krankheit mit möglichst hoher Sicherheit ausgeschlossen werden soll. Ein Beispiel aus der Buiatrik: Systolische Herzgeräusche kommen beim Rind relativ häufig vor und beruhen nicht immer auf Endokarditis. Bei Endokarditis treten fast stets Veränderungen im Eiweißgehalt des Plasmas auf, die mit dem Glutartest erfasst werden können. Sozusagen als Kehrseite der Medaille können massive entzündliche Prozesse wie Endokarditis weitgehend ausgeschlossen werden, wenn der Glutartest *negativ* verläuft.

Tests mit hoher diagnostischer Spezifität sind notwendig, wenn rasch einige Fälle einer Krankheit gefunden werden sollen, oder wenn das Ergebnis des Tests schwerwiegende Folgen (etwa im Form einer Operation mit erheblichem Risiko) nach sich zieht.

2.3.3-2: Man könnte auf den Gedanken kommen, dem Münzwurf einen prädiktiven Wert von 0,5 zuzumessen. Das ist in der Literatur auch schon geschehen. Aber das ist nicht korrekt.

	Mysteriose	**keine Mysteriose**	
Zahl (Test positiv)	a = 0,5*P	b = 0,5*(1-P)	0,5*P + 0,5*(1-P)
Kopf (Test negativ)	c = 0,5*P	d = 0,5*(1-P)	0,5*P + 0,5*(1-P)
	P	1-P	1

Mysteriose habe die Prävalenz P. Je in der Hälfte der Fälle erscheint beim Münzwurf Zahl (Test positiv; Feld a) oder Kopf (Test negativ; Feld c). Der Rest der Population ist 1-P. Auch hier fällt der Test in je der Hälfte positiv (Feld b) oder negativ (Feld d) aus. Definitionsgemäß gibt der prädiktive Wert (des positiven Testausfalls) den Anteil der Individuen mit der gesuchten Krankheit unter denen an, welche im Test positiv reagieren. Hier also:

$$pW(T+) = 0,5*P/[0,5*P + 0,5*(1-P)]$$

Diesen Ausdruck kann man umformen:

$$pW(T+) = 0,5*P/\ 0,5[P + (1-P)]$$
$$pW(T+) = P/(P + 1 – P)$$
$$pW(T+) = P$$

Das bedeutet, dass der prädiktive Wert des (positiven Ausfalls des) „Münzwurf-Tests“ gleich der Prävalenz (also der Apriori-Wahrscheinlichkeit = Wahrscheinlichkeit vor Durchführung des Tests) der fraglichen Krankheit ist. Mit anderen Worten, der Test liefert keinerlei Information, was beruhigend zu wissen ist.

Ein Test mit einem prädiktiven Wert von 0,5 wäre in vielen Situationen (in denen die Apriori-Wahrscheinlichkeit oder Prävalenz von Krankheiten ja viel niedriger ist) sehr hilfreich.

2.3.4-1: Die diagnostische Sensitivität von Symptom S1 im Hinblick auf Krankheit K entspricht definitionsgemäß dem Anteil der Individuen mit Krankheit K, welche das Symptom zeigen, hier also 8/24 = 0,33. Die diagnostische Spezifität ist der Anteil der Individuen ohne die Krankheit K (hier 76), welche in einem Test negativ reagieren oder ein Symptom nicht zeigen (hier 44), also 44/76 = 0,58. Für Symptom S2 sind die entsprechenden Werte: dSe = 18/24 = 0,75 und dSp = 64/76 = 0,84.

2.3.4-2: Bei disjunktivem Positivitätskriterium werden alle positiven Testausfälle als insgesamt positives Ergebnis gewertet. Die diagnostische Sensitivität der Kombination von S1 und S2 wäre hier also (8+12)/24 = 0,83. Sie liegt damit höher als die der beiden Symptome einzeln. Die diagnostische Spezifität wäre bei [76-(32+6)]/76 = 38/76 = 0,5. Sie ist damit geringer als die dSp der beiden Symptome einzeln. Bei Verwendung des konjunktiven Positivitätskriteriums werden nur übereinstimmend positive Ergebnisse als insgesamt positives Ergebnis gewertet. Im gezeigten Beispiel sind das sechs Individuen mit und sechs Individuen ohne Krankheit K. Damit ergibt sich eine dSe von 6/24 = 0,25. Sie ist geringer als diejenige der beiden Symptome einzeln. Dafür ist die dSp mit 70/76 = 0,92 höher als diejenige der beiden Symptome einzeln.

2.3.4-3: Bei Anwendung des disjunktiven Positivitätskriteriums steigt die Sensitivität. Dies ist anzustreben bei hoher Apriori-Wahrscheinlichkeit der in Frage stehenden Krankheit und/oder hohen Kosten für falsch negative Zuordnungen. Bei Anwendung des konjunktiven Positivitätskriteriums steigt die Spezifität. Dies ist anzustreben bei niedriger Apriori-

Wahrscheinlichkeit der in Frage stehenden Krankheit und/oder hohen Kosten für falsch positive Zuordnungen.

2.6-1: I entspricht Feld d (richtig negativ), II entspricht Feld c (falsch negativ), III entspricht Feld b (falsch positiv), und IV entspricht Feld a (richtig positiv).

2.6-2: Eine Verschiebung des Trennwertes nach rechts (also in Richtung der Lage der Werte kranker Individuen) führt zu einer Erniedrigung der diagnostischen Sensitivität (es wird ein immer kleinerer Anteil der Individuen mit der Krankheit K als „positiv“ eingestuft) und zu einer Erhöhung der diagnostischen Spezifität (es wird ein immer größerer Anteil der Individuen ohne die Krankheit K als „negativ“ eingestuft). Eine Verschiebung des Trennwertes nach links hat die umgekehrten Auswirkungen.

2.6-3: Die diagnostische Spezifität wurde erhöht, indem der Trennwert höher eingestellt wurde. Damit verbunden ist eine Senkung der diagnostischen Sensitivität.

2.7-1: Die Winkelhalbierende gibt die "Information" an, die durch den Zufall geliefert wird, entspricht also der Sensitivität und Spezifität des Münzwurfs.

2.7-2: Test B ist der bessere. Bei jeder gegebenen Sensitivität ist seine Spezifität höher, und bei jeder gegebenen Spezifität ist seine Sensitivität höher als die des Tests A.

2.7-3: Die „Kurve“ würde aus dem Punkt links oben im Koordinatensystem bestehen, bei dem sowohl Sensitivität als auch Spezifität den Wert 1 annehmen (1-Spezifität = 0).

2.7-4: Das "Trennungsvermögen", also die Fähigkeit, pathologische Phänomene zu erkennen, war bei den teilnehmenden Röntgenologen etwa gleich gut. Sie hatten aber bewusst oder unbewusst eine unterschiedliche Einstellung, indem sie eine der beiden möglichen Fehlerarten (falsch positiv oder falsch negativ) vorgezogen haben. Anders ausgedrückt: sie haben verschiedene Trennwerte verwendet.

2.9-1: p(K) entspricht der Prävalenz P; p(S/K) entspricht dem Anteil der Individuen mit richtig positivem Testausfall, also dSe. p(S) gibt den Anteil derjenigen Individuen an der fraglichen Population an, welche das Symptom S zeigen, also Individuen mit richtig positivem oder falsch positivem Testausfall. Erstere sind wieder gleich P*dSe, wie oben ausgeführt, letztere entspricht dem Produkt (1-P)*(1-dSp).

p (K+/T+) = pK * p(T+/K+) / pS+ pS+ = RP + FP

= Präv * dSe/(RP + FP)

= (Präv * dSe)/[(Präv * Empf) + (1 - Präv)*(1 - dSp)]

	Krankheit K vorhanden (K+)	Krankheit K nicht vorhanden (K-)
Test positiv (S+)	RP	FP
Test negativ (S-)	FN	RN
	RP + FN = Präv	FP + RN = 1 - Präv

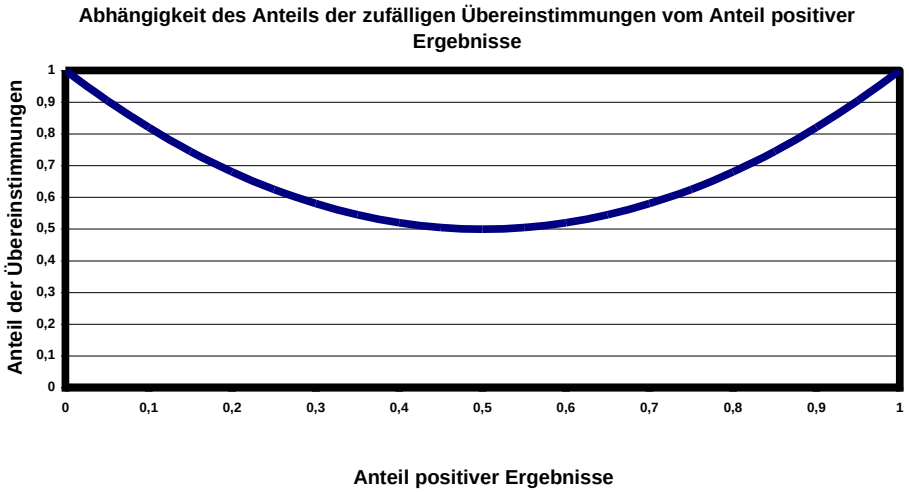
RP = Präv * dSe
RN = (1 - Präv) * dSp
FN = Präv * (1 - dSe)
FP = (1 - Präv) * (1 - dSp)

Allgemeine Formel für den pV T+

pVT+ = RP/(RP + FP)

pVT+ = (Präv * dSe)/[(Präv * dSe) + (1 - Präv)*(1 - dSp)]

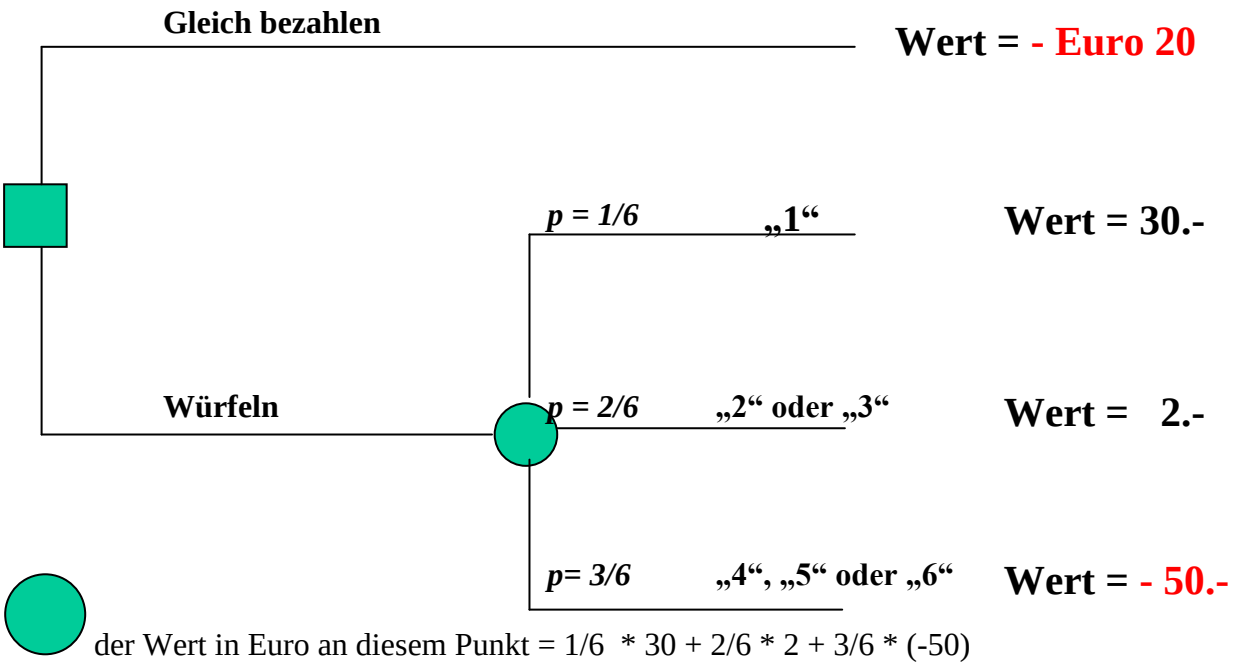
2.11-1: Die zufällig zu erwartende Übereinstimmung zwischen zwei Untersuchern (oder Tests) ist am kleinsten, wenn beide in je der Hälfte der Fälle ein positives (oder negatives) Ergebnis liefern. Die folgende Abbildung soll das illustrieren.



2.11-2: 0,42

3.2-1: Nein, denn die Ereignisse „Besserung“ und „Heilung“ sind nicht klar voneinander unterscheidbar.

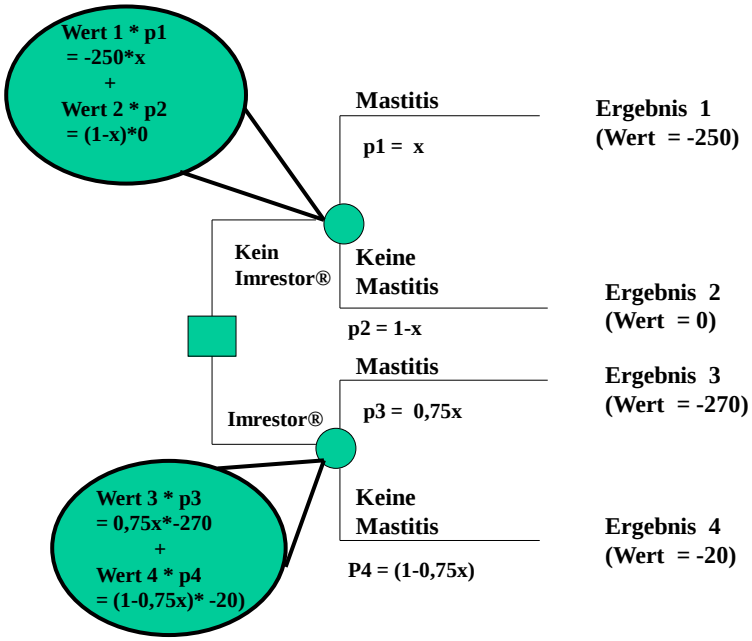
3.2-2



der Wert in Euro an diesem Punkt = 1/6 * 30 + 2/6 * 2 + 3/6 * (-50)
= - 19,33

Die Kosten beim Würfelt betragen „nur“ -19,33 Euro. Dieser Wert gilt im Mittel, wenn sehr viele Kunden würfeln. Dann sparen die Kunden im Mittel 67 Cent, obwohl einzelne mit 50 Euro einen sehr hohen Eintrittspreis bezahlen.

3.2.3: Hier ist der (sehr einfache) Entscheidungsbaum:



Der Einsatz von Imrestor® lohnt sich, wenn die zu erwartende Inzidenz größer als x ist.

-250x = 0,75x*-270 + (1-0,75x)*-20
-62,5 x = - 20
x = 0,32

7. Literaturhinweise

Einige der ausgewählten Arbeiten und Werke sind schon recht „betagt“ – wie der „Seniorverfasser“.

Abel, U. (1993) Die Bewertung diagnostischer Tests. Hippokrates.

Bonita, R, Beaglehole, R, Kjellström, T (2013) Einführung in die Epidemiologie 3. Aufl.

Bryant, GD, Norman, GR. (1980): Expressions of probability: Words and numbers. New Engl. J. Med. 302, 411)

Doohoo, I., Martin, W., Stryhm, H. (2003) Veterinary epidemiological research. AVC Inc, Chalottetoen

Evans, AS (1976): Causation and disease: The Henle-Koch postulates revisited. Yale J. Biol. Med. 49, 175-195

Feinstein, AR. (1967) Clinical judgment. Krieger Publishing Company, New York

Feinstein, AR. (1985) Clinical epidemiology. The Architecture of clinical research. Saunders

Feinstein, AR (1987) Clinimetrics. Yale University Press, New Haven, London

Fletcher, RH., Fletcher, SW., Wagner, EH. (1988) Clinical epidemiology. The essentials. 2. Auflage Williams & Wilkins

Gordis, L (2009) Epidemiology 4. Auflage. Saunders

Greiner, M. (2003) Serodiagnostische Tests. Springer

Groß, R (1969) Medizinische Diagnostik – Grundlagen und Praxis. Springer

Knapp, RG., Miller III, MC. (Clinical epidemiology and biostatistics. Harwal Publishing Company, Malvern

Kreienbrock, L., Pigeot, I., Ahrens, W. (2012) Epidemiologische Methoden. 5. Auflage Springer Spektrum

Last, JM. (Hrsg.) (1988) A dictionary of epidemiology. 2. Auflage Oxford University Press

Martin, SW (????) Veterinary epidemiology) ???

Pfeiffer, DU (2010) Veterinary Epidemiology. An Introduction. Wiley-Blackwell

Razum, O., Breckenkamp, J., Brzoska, P. (2011) Epidemiologie für Dummies. 2. Auflage. Wiley VCH

Scadding, JG (1967) Lancet, 290: 877-882

Smith, RD (2005) Veterinary Clinical Epidemiology 3. Auflage. Taylor & Francis

Sox Jr., HC, Blatt, MA., Higgins, MC., Marton, KI. (1987) Medical decision making. Butterworths

Toogood (1980): What do we mean by "usually", Lancet 1980-1 10984

Vecchio, TJ. (1966) N Engl J Med; 274:1171-1173

Villaroel, A. (2015) Practical clinical epidemiology for the veterinarian. Wiley Blackwell

Weinstein, MC, Fineberg HV, Elstein, AS, Frazier, HS, Neuhauser, D, Neutra, RR, Mcneil, B (1980) Clinical decision analysis. Saunders

Wieland, W. (1975) Diagnose. Überlegungen zur Medizintheorie. Walter de Gruyter. Berlin, New York

8. Stichwortverzeichnis

Abfolge von Tests 22
Abort 59, 60
Altersgruppe 87
Aposteriori-Wahrscheinlichkeit 45
Apriori-Wahrscheinlichkeit 15, 19, 20, 45
Aristoteles 79
Assoziation 87, 88
Atemwegserkrankung 31
Attributable Risk 65
Auswahlstrategie 63
BAYES 45
Befunde 9
Behandlungen 87
Beschwerden 9
Bias 42
Blauzungenkrankheit 56
Blickdiagnosen 50, 51
BOOLEsche Algebra 28
bovine neonatale Panzytopenie 6
bovine Virusdiarrhoe 53
Brucellose 95
Bullenmäster 53
Chancenverhältnis 87
Chi-Quadrat-Test 40, 47, 67, 88
Confounder 90
CONSULTANT 28
cutoff 32, 37, 40
Cutoff 39
Definitionskriterium 43
Diagnose 5, 9, 42, 93
Diagnostik 5, 14, 28, 31, 37, 43, 50, 51, 53, 94
diagnostische Sensitivität 10
diagnostische Spezifität 11
diagnostisches Repertoire 94
dichotom 32
Differentialdiagnostik 31
disjunktives Positivitätskriterium 22
diskrepante Ergebnisse 21
Diskriminanzanalyse 20
EBP 75
Effektivität 12, 39
Eindeutigkeit 7, 44
EKG 37
ELISA 26
empirische Wahrscheinlichkeit 84
Endokarditis 32, 95
Entscheidungsbaum 69, 75
Entscheidungsknoten 70
epidemiologische Kurve 63
Erfahrung 51, 79, 93
Euterentzündung 50
EVANS 80
Exhaustionsmethode 50
Exposition 65, 81
Faktor 86
Falldefinition 42
Fall-Kontroll-Studie 88
Fehldiagnosen 50, 93
Formel 24
Fremdkörperperitonitis 11
Futterkomponenten 87
GAUSS-Verteilung 61
Gesundheit 31
glomeruläre Filtrationsrate 42
Glutartest 32, 95
Goldstandard 37, 39, 42, 67
Good Clinical Practice 67
GÖTZE 50
GROSS 6
Hämoglobinkonzentration 60
HENLE-KOCH 80
hinreichende Wahrscheinlichkeit 85
Hyänenkrankheit 6
Hypokalzämie 32
Hypothese 50
Impfkosten 76
Impfungen 87
Inkubationszeit 80
Inulin-Clearance 42
Inzidenz 75, 80, 81, 82
Inzidenzdichte 83
Irrtumswahrscheinlichkeit 59
Jungmastbullen 31
Kälbermastbetrieb 60
KANT 79
Kappa 48, 49
Kappa-Statistik 46
Kasuistik 27
Kausalität 79
KBR 26
Klassifikation 7
Kliniker 32, 43, 50, 69
Klinische Entscheidungsanalyse 69

klinische Untersuchung 17
Kohortenstudie 89
Kombination 19, 26
Kombination von Tests 19
Konfidenzintervall 60, 88
konjunktives Positivitätskriterium 22, 24
Kontrollgruppe 16, 17, 67
Koronarinsuffizienz 37
Korrelation 21, 24, 25, 26
Krankheit 13, 16, 28, 31, 42, 50, 51, 81
Krankheitsdefinition 44
kumulative Inzidenz 82
Labmagenverlagerung 20
Labordiagnostik 18, 31, 32
Letalität 67
Likelihood ratio 18
LINNÉ 7
Listeriose 93
logische Verzweigung 51
logische Wahrscheinlichkeit 84
logistische Regression 88
Lotto 84
Medien 66
Merzung 55
Meta-Analysen 73
metabolische Profiltests 62
Milchmastkälber 60
Mobilfunkfelder 91
Münzwurf 96
Mustererkennung 51
Mysteriose 86
NAEGELI 5
Nierenfunktionsdiagnostik 42
nominalistischer Krankheitsbegriffes 5
Normalverteilung 80
Normalwerte 31
Nosographie 27
nosologische Einheit 9
nosologischen Einheit 5
Nullzähler 59
Numbers needed to treat 67
Nutztiermedizin 74
Ockhams Rasiermesser 52
Odds Ratio 65, 87
ontologischer Krankheitsbegriff 5
Outbreak Investigation 63
parallel 21
Paratuberkulose 25, 63
Pathologie 43
persistierende Infektion 53
Population 9, 15, 17, 53, 55, 82, 86
Populationen 37
prädiktiver Wert 13, 14, 15, 17
Prävalenz 15, 17, 19, 22, 35, 45, 53, 55, 82
Praxisnähe 7
Präzision 18, 55
problemorientierte Diagnostik 52
Propädeutik 50
Prophylaxe 79
Punkt-Exposition 64
Querschnittsuntersuchung 86, 88
Reagenten 35
Referenzintervall 30
Referenzintervalle 31
Referenzmethode 42
Reihenfolge 21
relatives Risiko 90
Relatives Risiko 65
repräsentative Stichprobe 54
Richtigkeit 12, 18, 40, 50
Rinderleukose 55
Rindertuberkulose 35, 36, 95
Risiko 53, 82, 83
ROC-Kurve 30, 37, 39
Roulette 84
rule-outs 52
SCADDING 6
scheinbare Prävalenz 18
schützende Wirkung 65
Schwindsucht 43
Schwingauskultation 46
Sensitivität 10, 11, 12, 14, 15, 27, 31, 32, 37, 53, 93, 95
Sensitivitätsanalyse 69, 74, 77
seriell 21
Seuchenfreiheit 63
Spannweite 35
Spezifität 11, 12, 14, 15, 32, 37, 53
Stallabteilung 87
Standardabweichung 48
statistische Signifikanz 40
Stichprobe 12, 40, 55, 59, 88
subjektive Wahrscheinlichkeit 84
SUPPES 79
SYDENHAM 5, 7
Symptom 9, 23
Systematik 7
Tabellenkalkulation 77
Test 11, 13, 16, 17, 26

Tests 9, 11, 12, 20, 21, 42, 53
Therapie 67
Thiaminmangel 32
Tierschutz-Nutztierhaltungsverordnung 60
Tollwut 93
Tränke 87
Trennwerte 32, 37
Tuberkel 43
Tuberkulintest 35
Tuberkulose 43
Übereinstimmung von Testergebnissen 46
Ursache 6, 79, 81, 88
Validität 12, 16, 53
VECCHIO 13
Vereinfachung 29, 32
Vertrauensbereich 48, 56, 57, 58, 60, 88
Vierfeldertafel 9, 10, 13, 14, 33, 65, 82, 86, 87, 90
Vollständigkeit 7, 44
Vorbericht 20, 52, 88
Vor-Test-Wahrscheinlichkeit 15
Wahrnehmung 92
Wahrscheinlichkeit 25, 56, 60, 84
WHO 31
Winkelhalbierende 38
Youden-Index 40
Zufallsknoten 70
Zweckmäßigkeit 7