Modellierung und Statistik in der Medizin – Risiken und Entscheidungen unter Unsicherheit

MANFRED BOROVCNIK, KLAGENFURT

Informationen zu Gesundheitsfragen werden im Allgemeinen durch statistische Methoden begründet; Entscheidungen sind idealerweise evidenzbasiert und durch Wahrscheinlichkeitsmodelle gestützt. Um die Rationalität im Umgang mit solchen Informationen zu erhöhen, befürworten Experten die Methodik der empirischen Forschung, welche komplexe mathematische Konzepte voraussetzt, die kaum verstanden werden. Im Mittelpunkt des Aufsatzes steht die Statistik in der Medizin und die Tatsache, dass es nicht zuletzt mit dem zunehmenden Informationszugang über das Internet Vor- und Nachteile gibt. Diejenigen, die an der Verbesserung der Entscheidungsqualität interessiert sind, konzentrieren sich auf Ansätze und Strategien, um die vorgegebenen Methoden (und damit die daraus resultierenden Informationen) besser zu verstehen. Wir werden den Rahmen und verschiedene Ansätze zur Förderung von Risikokompetenz skizzieren. Themen in Einzelnen: Komponenten von Situationen unter Unsicherheit, Risikomanagement in Gesundheitsfragen, Statistische Methoden in der Medizin. Die Überlegungen werden auch anhand von konkreten Fallstudien illustriert.

1. "Methoden" der Risiko-Literalität

Wir gehen zuerst auf die Definition von Risiko und verwandten Konzepten ein und zeigen, dass bei kleinen Wahrscheinlichkeiten eine Schätzung durch Daten praktisch ausgeschlossen ist, was natürlich auch eine Deutung dieser Wahrscheinlichkeiten als relative Häufigkeiten ernsthaft in Frage stellt. Sodann zeigen wir, wie man bedingte Wahrscheinlichkeiten in erwartete Häufigkeiten umrechnen kann, und wie man aus den entsprechenden Tabellen und verschiedensten Diagrammen die erwarteten Häufigkeiten und damit die Situation im Kontext besser beurteilen kann. An Diagrammen behandeln wir neben dem Baumdiagramm auch das Mosaikdiagramm und das Ikon-Diagramm. Zu Risiko und dem Verhältnis zwischen Risiko und Wahrscheinlichkeit findet man mehr in Borovcnik (2015, 2016a) sowie Borovcnik und Kapadia (2018), zwei grundsätzliche Aufsätze des Autors zum Thema Risiko und Modellierung in der Medizin sind Borovcnik (2019a, b). Didaktisch reizvoll ist auch der Ansatz von Borovcnik (2016b), einem fiktiven Dialog von Betroffenen zuzuhören, der sich mit der Auseinandersetzung mit einer positiven Krebsdiagnose befasst.

1.1 Risiko: Definition und verwandte Konzepte

Risiko und Entscheidungen

Unter Risiko verstehen wir eine Situation mit inhärenter Ungewissheit über (zukünftige) Ereignisse, die zu einem Impakt (Schäden oder Vorteile) führen. Manchmal wird der Erwartungswert des Impakts genommen, um verschiedene Optionen miteinander zu vergleichen. Oft muss eine Entscheidung zwischen festen Optionen getroffen werden.

Risiko wird in inkonsistenter Weise verwendet

Es werden häufig nur die extremen Situationen ins Kalkül gezogen, eine Abwägung dazwischen – was einer rationalen Vorgehensweise entsprechen würde – wird selten durchgeführt: Nur mit Bezug auf die Wahrscheinlichkeit des unerwünschten Ereignisses, ohne Konsequenzen (Impakt) in Betracht zu ziehen. Oder das andere Extrem: Nur die Konsequenzen, ohne die Wahrscheinlichkeiten zu beachten.

Die Entscheidung kann unterschiedliche Auswirkungen haben

Die Auswirkungen einer Entscheidung können einzelne Personen, zwei oder mehrere Personen, oder Menschen und eine Institution betreffen.

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 54, 2022, S. 1 - 16.

Versuche, Risiko zu definieren

(1) Ein unerwünschtes Ereignis, das eintreten oder ausbleiben kann.
(2) Die Ursache eines unerwünschten Ereignisses, das eintreten oder ...
(3) Die Wahrscheinlichkeit eines unerwünschten Ereignisses, das eintreten oder …
(4) Der Erwartungswert eines unerwünschten Ereignisses, das eintreten oder …
(5) Die Tatsache, dass eine Entscheidung unter Bedingungen getroffen wird, deren Wahrscheinlichkeiten bekannt (und daher nicht unbekannt) sind.

Eine detaillierte Kommentierung von Definitionsversuchen zu Risiko ist Borovcnik (2015) zu entnehmen. Die Wahrnehmung von Risiken ist eng verbunden mit der von Wahrscheinlichkeiten. Die Kombination von schwerwiegenden Folgen und niedriger Wahrscheinlichkeit verzerrt die Wahrnehmung von Risiken enorm. Definition (4) kommt am nächsten zur rationalen Entscheidungstheorie (siehe etwa Mongin, 1997).

Mit Risiko verwandte Konzepte

In der beurteilenden Statistik spricht man vom Risiko falscher Entscheidungen beim statistischen Test. Als Fehler 1. Art (alpha-Fehler) wird die Wahrscheinlichkeit bezeichnet, die sogenannte Nullhypothese irrtümlich abzulehnen (wenn sie also nicht zutrifft), als Fehler 2. Art wird die Wahrscheinlichkeit bezeichnet, die Nullhypothese zu Unrecht (fälschlich) nicht zu verwerfen (wenn sie also nicht zutrifft). Erst die erläuternden Ergänzungen in Klammern zeigen auf, dass es sich bei diesen Risiken um keine absoluten Wahrscheinlichkeiten handelt sondern um bedingte Wahrscheinlichkeiten. Dass man aus diesen Risiken erst durch Kenntnis weiterer Wahrscheinlichkeiten (der sogenannten a priori-Wahrscheinlichkeit der Nullhypothese) eine Risikobeurteilung global treffen kann, wird einem zunächst gar nicht bewusst. Die Sprache lässt zudem die Bedingung unter den Tisch fallen und so entsteht – fälschlicherweise – der Eindruck, dass es sich hierbei um einfache (also nicht bedingte) Wahrscheinlichkeiten handelt. Daraus resultieren viele Fehlinterpretationen bei der medizinischen Diagnose von Krankheiten, die man auch als statistischen Test interpretieren kann (siehe etwa Borovcnik, 2019b, sowie Borovcnik & Kapadia, 2018).

- *Risk und Hazard*. Hazard als "Ursache" eines unerwünschten Ereignisses.
- Risk und Uncertainty. Knight (1921) sieht die Differenz darin, ob die Wahrscheinlichkeiten bekannt sind oder nicht.
- Risiko und Nutzen in der theoretischen Ökonomie. Maximierung des erwarteten Nutzens u. Risiko-Aversion als u´´/u´
- System-analytischer Ansatz zum Untersuchen von Situationen mit Risiko. Welche Konstituenten? Wer? Typ des Risikos? Information? Welche Situation (Vorsorge oder erzwungene Entscheidung)?

1.2 Das Problem kleiner Risiken und kleiner Wahrscheinlichkeiten

Für kleine Wahrscheinlichkeiten ist eine Interpretation durch Häufigkeiten obsolet: Wir haben einfach keine Daten dafür. Alle Wahrscheinlichkeitswerte stammen aus Modellen auf der Basis von Annahmen. Eine Häufigkeitsdeutung dieser Wahrscheinlichkeiten ist artifiziell und kaum hilfreich.

Beispiel der Bovinen spongiformen Enzephalopathie (BSE, Dubben & Beck-Bornholdt, 2010)

Alle positiven Rinder in Deutschland könnten sehr wohl falsch positiv gewesen sein. Sensitivität (0.99) und Spezifität (0.997) wurden auf der Basis von 300 bzw. 1000 Rindern unter reinen Laborbedingungen geschätzt. Ein Simulationsszenario der Schätzung einer ansonsten unbekannten Wahrscheinlichkeit p von 0.0001 (viel größer als die Inzidenz von BSE) durch 10000 zufällige (!) Daten zeigt, dass sehr große Fehler möglich sind (Abb. 1). Fazit: Es gibt keine Möglichkeit, kleine Wahrscheinlichkeiten zuverlässig zu schätzen!

		Schätzung von p	Relativer Fehler der Schätzung	Anzahl	%	Wahrschein-lichkeit
	0	0.0000	-100%	34	34.0%	0.3679
	1	0.0001	0%	40	40.0%	0.3679
	2	0.0002	100%	18	18.0%	0.1839
	3	0.0003	200%	3	3.0%	0.0613
	4	0.0004	300%	3	3.0%	0.0153
	5	0.0005	400%	2	2.0%	0.0031
	6	0.0006	500%	0	0.0%	0.0005
	7	0.0007	600%	0	0.0%	7.29E-05
	8	0.0008	700%	0	0.0%	9.11E-06
	9	0.0009	800%	0	0.0%	1.01E-06
	10	0.0010	900%	0	0.0%	1.01E-07
				0		
				100	100.0%	1.0000

Das ist der **wahre Wert von p**, den man aus den 10 000 Daten zu schätzen hat.

Die Schätzung einer unbekannten Wahrscheinlichkeit bleibt auch bei enorm vielen Daten sehr ungenau & unsicher.

Abb. 1: Simulationsszenario mit 100 Schätzungen einer sehr kleinen, unbekannten Wahrscheinlichkeit mit 10000 Daten

Szenario: Ein Bestand an „Parkplätzen“, die zufällig von Ticketbesitzern in Anspruch genommen werden. Abhängig von den der Wahrscheinlichkeit p der Nutzung kann man mehr Tickets verkaufen als Parkplätze vorhanden sind. Das birgt ein Risiko der Überbuchung. Um dieses Risiko zu verkleinern, muss man auf den finanziellen Vorteil aus einer Überbuchung gänzlich verzichten.

Szenario einer Prüfung mit Single-Choice-Items (ja-nein): Wir nehmen an, dass ein Prüfling die Items unabhängig voneinander mit derselben Erfolgswahrscheinlichkeit p löst. Wegen der 50%-Hürde besteht ein Risiko, in der Prüfung zu versagen. Wir modellieren n Items (n = 10, 30, 100) und spielen verschiedene Werte von p für das “Potential“ durch. Man orientiert sich rasch interaktiv an einer graphischen Darstellung der simulierten Ergebnisse, dass man weit über 50% Lösungskapazität für ein einzelnes Item haben muss, damit man das Risiko durchzufallen entsprechend klein hält. So besteht etwa bei einer Lösungskapazität von 66% für einzelne Items noch immer ein Risiko von 2.3% durchzufallen. Will man dieses auch noch verkleinern, so steigt die erforderliche Kapazität extrem an.

Fazit: Um Risiken zu verringern, bedarf es enormer Anstrengungen. Wir konnten in den letzten Jahren sehen, wie schwierig sich öffentliche Maßnahmen im Zusammenhang mit der Pandemie gestalteten, weil man Risiken immer weiter verkleinern wollte. Wahrscheinlichkeiten, insbesondere kleine Wahrscheinlichkeiten können nicht numerisch interpretiert werden, sie können höchstens verglichen werden im Sinne von „ist kleiner“. Kleine Wahrscheinlichkeiten sind damit höchstens ordinal. Historisch gesehen hat man von einer moralischen Wahrscheinlichkeit gesprochen und diskutiert, ob man nicht alle Wahrscheinlichkeiten, die den Schwellenwert der moralischen Wahrscheinlichkeit unterschreiten, mit Null gleichgesetzt werden sollen. Als Werte waren 10^{-4} im Gespräch. Heute untersuchen wir routinemäßig Risiken, die weit unter solchen Schwellenwerten liegen.

1.3 Umrechnung bedingter Wahrscheinlichkeiten in Erwartungswerte

Szenario: Für Frauen zwischen 40 und 50 sind folgende Risiken für das Mammographie-Screening (eine präventive Untersuchung ohne Vorliegen spezifischer Symptome) gegeben. Als Prävalenz der Krankheit nehmen wir in Einklang mit Daten für die Wahrscheinlichkeit für Brustkrebs einen Wert von 0.8 % an. Für die Zuverlässigkeiten der Diagnoseprozedur gelte

- Wenn eine Frau Brustkrebs hat, ergibt die Mammographie ein positives Resultat mit einer Zuverlässigkeit (einer bedingten Wahrscheinlichkeit) von 90 %.

- Wenn eine Frau keinen Brustkrebs hat, dann besteht ein Risiko von 7 % (eine bedingte Wahrscheinlichkeit) für eine positive Mammographie.

Man braucht die Bayes-Formel, um die wirklich interessierende Wahrscheinlichkeit zu berechnen. Wir nehmen and, dass eine Frau einen positiven Mammographie-Befund hat. Wie groß ist die Wahrscheinlichkeit, dass sie tatsächlich Brustkrebs hat?

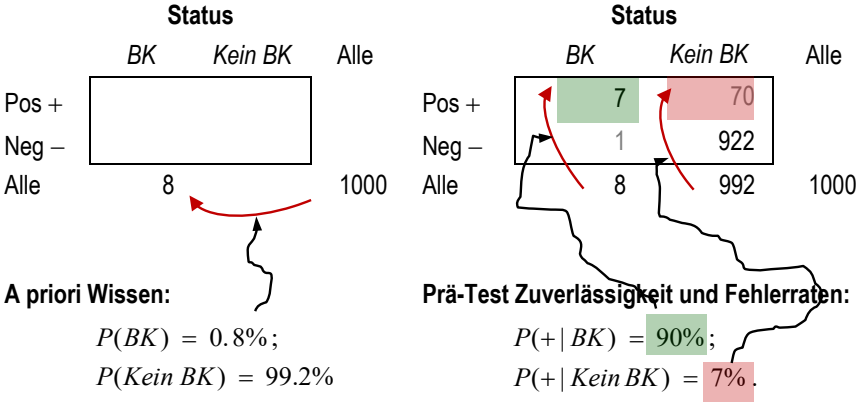

Abb. 2: Bedingte und andere Wahrscheinlichkeiten werden in erwartete Häufigkeiten umgerechnet.

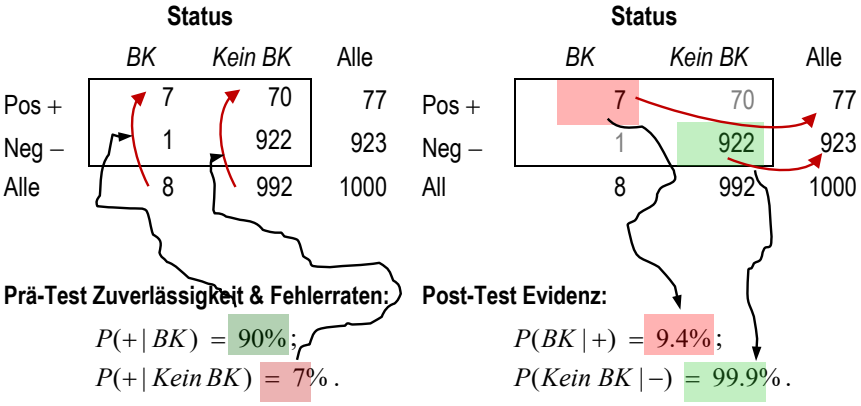

Abb. 3: Interpretation dieser erwarteten Häufigkeiten als Wahrscheinlichkeiten

Die gefragten Wahrscheinlichkeiten sind aus dieser Tabelle leicht ablesbar (siehe auch Batanero & Borovcnik, 2016). Gigerenzer (2002) benennt diese als natürliche Häufigkeiten. Spiegelhalter und Gage (2015) benutzen dieselbe Technik bleibt aber bei erwarteten Häufigkeiten. Spiegelhalter (2012) versucht, diese Erwartungswerte bei sehr kleinen Wahrscheinlichkeiten in sogenannte „microlives“ umzurechnen, um sie intuitiv erfassbar zu machen.

1.4 Tabellen und graphische Darstellungen von Daten mit erwarteten Häufigkeiten

Wir zeigen die Kraft der Darstellungen, indem wir die erwarteten Häufigkeiten in einer Kontingenztabelle anordnen und in einem Baumdiagramm oder einem Mosaik- bzw. Ikon-Diagramm darstellen.

Anordnen der erwarteten Häufigkeiten in einer Kontingenztabelle

Die Spalten in Abb. 3 entsprechen der gegebenen Information, nämlich den bedingten Wahrscheinlichkeiten für die Güte der Diagnose. Aus der ersten Zeile liest man die gesuchte Wahrscheinlichkeit ab, dass eine Frau mit positivem Mammogramm tatsächlich Brustkrebs hat: 7/77 = 1/11 ≈ 9.4%.

Im üblichen Format muss man erst lernen, wie die Information richtig zu lesen ist. Und die Bayes-Formel anwenden, um die gesuchte Wahrscheinlichkeit zu berechnen. Viele Formate in wissenschaftlicher Kommunikation verbergen die inhärente Information mehr, als dass sie sie hervorheben.

Anordnen der erwarteten Häufigkeiten in einem Baumdiagramm

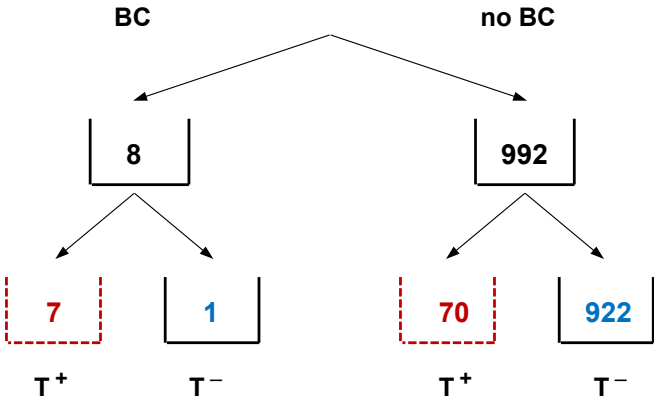

Abb. 4: Baumdiagramm für die Brustkrebsdaten

Die positiven Diagnosen sind über zwei Sorten von Urnen verteilt. Die eine wird gefüllt durch jene, die Brustkrebs haben, die andere von jenen, die keinen Brustkrebs haben. Wir prüfen, woher die Befüllung kommt und können damit die gesuchte bedingte Wahrscheinlichkeit bestimmen (rote Urnen): 7 von (7+70).

Graphische Darstellung der erwarteten Häufigkeiten in einem Mosaik-Diagramm

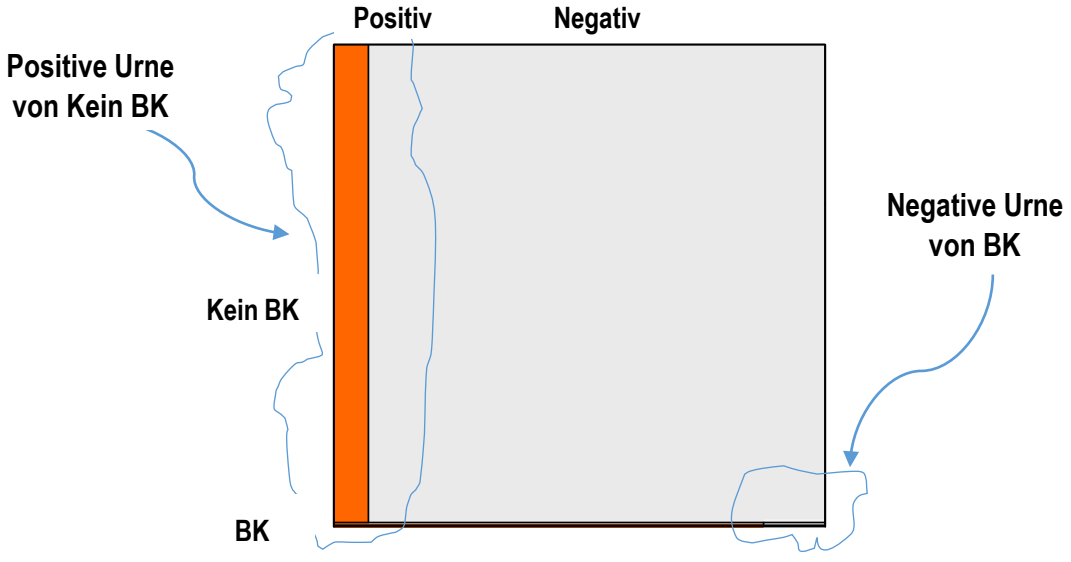

Abb. 5: Mosaik-Diagramm für die Brustkrebsdaten

Man teilt erst die Höhe des Einheitsquadrats nach der Prävalenz der Krankheit, sodann teilt man in den beiden horizontalen Streifen nach den jeweiligen bedingten Wahrscheinlichkeiten für eine positive Diagnose (siehe auch Batanero & Borovcnik, 2016). Die Fläche repräsentiert gleichzeitig absolute Häufigkeiten und Prozentsätze. Jetzt kann man alle Verhältnisse direkt als Flächenproportionen sehen. Man „sieht“, dass sich die Masse der positiven Diagnosen (L-Form in orange) überwiegend auf die Frauen der Gruppe *Kein BK* erstreckt (10/11; mehr als 90% geschätzt). Das entspricht der annähernd 10% Wahrscheinlichkeit, tatsächlich Brustkrebs zu haben, wenn man eine positive Diagnose hat.

Graphische Darstellung der erwarteten Häufigkeiten in einem Ikon-Diagramm

Jede Person wird durch ein Ikon repräsentiert, das gemäß ihrer Attribute gefärbt ist: Je dunkler ein Ikon, desto „schwerer“ der Fall. Das lässt allenfalls auch einen Rückschluss auf die Individuen zu, wenn Auffälligkeiten erklärt werden sollen. Der Eindruck ist gemischt visuell und quantitativ. Wir zeigen die Stärke dieses Diagramms anhand von Daten aus einer Prostatakrebsstudie (vgl. Spiegelhalter, 2014).

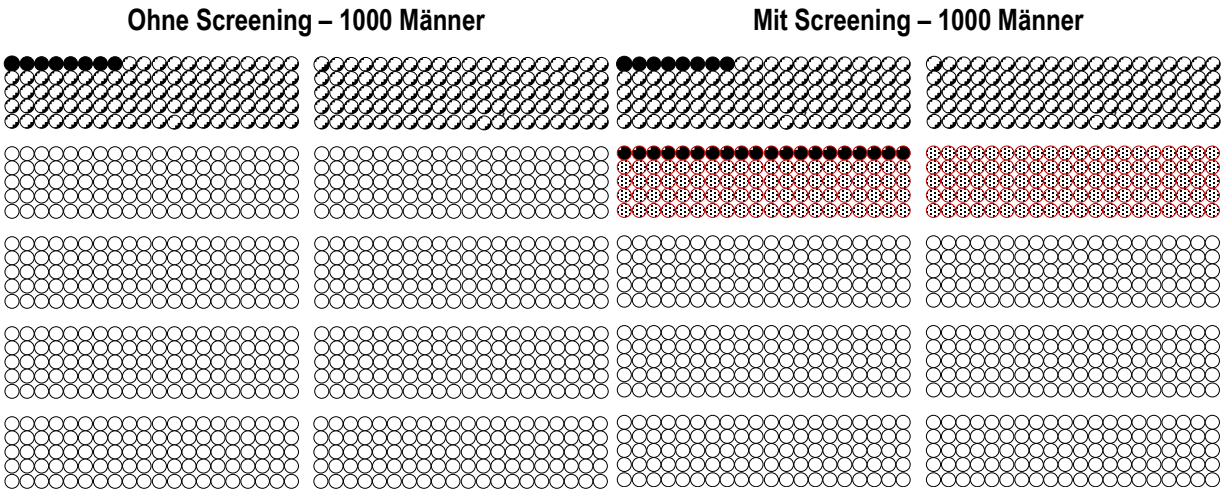

Abb. 6: Ikon-Diagramm illustriert die erwarteten Ausgänge für das Prostata-Screening –
verglichen werden zwei Gruppen von je 1000 Männern (50+), eine ohne und eine mit Screening im Programm über 10 Jahre.

Der Eindruck schärft sich noch, wenn man das Diagramm in Farbe darstellt. Hier bietet sich eine Graduierung von tief-rot nach grün. Tabelle 1 zur Metastudie sagt dasselbe aus, verblasst jedoch gegen die Kraft des Visuellen.

Tab. 1: Erwartete Anzahlen einer 10-Jahres-Screening Gruppe verglichen mit einer Gruppe ohne Screening;
jeweils 1000 Männer (50+).

Screening			Kategorie des Outcome
Mit	Ohne	Ikon	
8	8		Verstorben an Prostata-Krebs
192	192		Verstorben aus anderen Gründen
–	20		Prostatakrebs-Diagnose & unnötige Behandlung
–	180		Ohne Krebs, aber Biopsie (falscher Alarm)
800	600		Ohne Beeinträchtigung, lebend
1000	1000		Alle

Die Metaanalysen von Koubenec (2000) und Sandbløm et al. (2011) kommen zu ähnlichen Einschätzungen von Screening-Programmen zur frühen Entdeckung von Brust- bzw. Prostata-Krebs. Gigerenzer äußert sich in einem Interview (Wewetzer, 2005) ähnlich. Zu denselben Schlüssen über den zweifelhaften Erfolg von langfristigen Screening-Programmen kommt auch Gigerenzer (2014), der die Ergebnisse von Gøtzsche und Nielsen (2011) zusammenfasst. Dabei sind natürlich die Folgen einer häufigen Strahlenbelastung noch gar nicht erfasst.

2. Statistische Methoden in der Medizin

Im Wesentlichen werden medizinische Erkenntnisse aus kontrollierten Experimenten gewonnen, bei welchen der Outcome einer Behandlung evaluiert wird, indem man einer Gruppe von Patienten das sogenannte Verum verabreicht und der anderen Gruppe ein Placebo oder die gegenwärtig als Gold-Standard geltende Behandlung gibt. (Letzteres, falls es aus ethischen Gründen unmöglich ist, den Patienten lediglich eine Scheinbehandlung – das Placebo – zukommen zu lassen.) Ausgewählt werden die Patienten, welche die experimentelle Behandlung erhalten, durch Zufall. Man spricht vom randomisierten, kontrollierten Experiment. Wirkung und jedwede Nebenwirkung werden über die Dauer des Experiments akribisch aufgezeichnet, möglichen Fehlentwicklungen wird durch laufendes Monitoring kaum Platz gegeben. Hier sei auch auf Borovcnik (2021) verwiesen.

2.1 Signifikanztests und das randomisierte kontrollierte Experiment

Der Signifikanztest

Die Auswertung der Daten erfolgt durch statistische Tests; von nicht-parametrischen Methoden wie dem Wilcoxon-Test, dem t-Test zum Vergleich von zwei Gruppen bis hin zur Varianzanalyse mit allen methodischen Feinheiten. Gemeinsam ist allen diesen Tests die Methode des reinen Signifikanztests. Der Signifikanztest kann keine Alternativhypothesen berücksichtigen und entspricht im weitesten Sinne einer Abschwächung des mathematischen Beweises durch Widerspruch. Der „Widerspruch" besteht hier in weicher Form, dass man sagen kann, wenn die Nullhypothese, dass kein Unterschied zwischen den Patienten der Behandlungs- und der Kontrollgruppe besteht, zutrifft, dann werden das beobachtete Ergebnis und noch extremere in höchstens 5% (1%) der Fälle auftreten.

Die Macht (Power) des Tests, die Wahrscheinlichkeit, dass der Test zur Ablehnung der Nullhypothese gelangt, wenn eigentlich eine ganz andere Hypothese zutrifft, kann man beim reinen Signifikanztest nach R.A. Fisher nicht bestimmen, auch nicht in dessen modernen Abwandlungen, die den *p*-Wert aus den Daten berechnen statt das Signifikanzniveau fix vorzugeben. Das war und ist bis zum heutigen Tag ein Kern der kritischen Methodendiskussion (siehe etwa Hubbard & Bayarri, 2003, Wasserstein & Lazar, 2016, und ASA, 2016). Auch dazu findet man – im medizinischen Kontext – mehr in Borovcnik (2021).

Der *p*-Wert kann grob so umschrieben werden:

$$p = \text{Wahrscheinlichkeit für „ein beobachtetes Resultat (oder extremere)“, wenn die Nullhypothese } H_0 \text{ zutrifft}$$

Wenn *p* kleiner ist als 5%, wird die Nullhypothese verworfen; *p* ist die Wahrscheinlichkeit für eine irrtümliche Ablehnung der Nullhypothese, d.h., der Test erbringt ein signifikantes Ergebnis, falls die Nullhypothese zutrifft:

$$p \text{ (Test signifikant } | \text{ } H_0 \text{ trifft zu)}.$$

Wir haben dann etwas beobachtet, das weniger als 5% Wahrscheinlichkeit hat, wenn H_0 zutrifft. Wir sehen uns daher berechtigt, die Nullhypothese abzulehnen.

Klar ist, dass die Formulierung, „ein beobachtetes Resultat (oder extremere)" schwammig ist. Sie ist aber auch keineswegs eine unbedingte Wahrscheinlichkeit, wie etwa die Sechs beim Werfen eines Würfels zu bekommen. Nein, sie ist eine bedingte Wahrscheinlichkeit, bedingt auf den Zustand, dass die Nullhypothese auch tatsächlich zutrifft.

Eine erste Übertragung der Situation im statistischen Test auf die Diagnose einer Erkrankung

Im Kontext der medizinischen Experimente entspricht der Nullhypothese „kein Unterschied“ zwischen den beiden Gruppen von Behandelten (Versuchsgruppe) und Nicht-Behandelten (Kontrollgruppe). Eine Ablehnung der Nullhypothese wird dann mit der Wirksamkeit der Behandlung gleichgesetzt.

Oder, die Nullhypothese bei der Diagnose bestimmter Krankheiten ist „der Patient hat diese spezielle Krankheit NICHT“. Eine Ablehnung hier kommt einer positiven Diagnose gleich, d.h., man entscheidet, dass der Patient diese Krankheit hat und handelt dementsprechend.

Das Signifikanzniveau eines Tests (oder auch der p-Wert) sind, wie schon festgestellt, bedingte und nicht absolute Wahrscheinlichkeiten. Besonders deutlich wird das im Kontext mit der Diagnose:

H_0: Patient hat diese bestimmte Krankheit K NICHT; H_1: Patient hat die Krankheit K

Das Signifikanzniveau α, bei dessen Unterschreitung wir die Diagnose positiv ausstellen und so handeln, als ob der Patient diese Krankheit hat, bezieht sich ja auf die Untergruppe der Krankheit und nicht auf alle. Der Kontext der Diagnose zeigt noch mehr auf, nämlich, dass wir eigentlich an einer ganz anderen bedingten Wahrscheinlichkeit interessiert sind, die aber meist unbekannt ist:

$$P(\text{Patient hat die Krankheit K wirklich} \mid \text{Diagnose positiv})$$

Im Kontext der Diagnose lassen sich die Begrifflichkeiten und deren Interpretierbarkeit noch besser abklären. Dazu im nächsten Abschnitt mehr. An dieser Stelle sei nur gesagt, dass die Qualität des Signifikanztests speziell darunter leidet, dass man keinerlei Betrachtungen miteinbezieht für den Fall, dass die Nullhypothese NICHT zutrifft. Wie sich das Verfahren einer Entscheidung basierend auf dem Signifikanztest entwickelt, wie man seine Qualität messen kann, wird also überhaupt gar nicht erfasst.

2.2 Medizinische Diagnose basierend auf Trennpunkten, um die Gruppen Gesund und Krank zu separieren

Speziell der Fall der Trennung zweier Gruppen (Gesunde vs. Kranke) entspricht weitgehend einem statistischen Test. Jetzt werden *zwei* Verteilungen (je nach Zustand) für ein bestimmtes Merkmal verglichen. Zwei Verteilungen und nicht viele, weil das die Situation vereinfacht, aber immer noch die wesentlichen Eigenschaften der Entscheidungssituation mit ihren potentiellen Fehlern widerspiegelt. Für die Medizin ist dann wesentlich, dass man die Trennung, die Diagnose, weitgehend optimiert, d.h., dass man Fehlzuordungen bzw. Fehldiagnosen so weit wie möglich vermeidet oder wenigstens geringhaltet. Hier sind zwei Fragestellungen wichtig: Welches Merkmal lässt eine gute Trennung, also eine gute Diagnose der Krankheit zu? Wie groß sind die Fehler einer falschen Zuordnung? Auch hierzu findet man in Borovcnik (2021) eine datenorientierte Analyse und eine begriffliche Aufarbeitung.

Fehlermöglichkeiten bei der Diagnose einer Krankheit

In Abbildung 7 zeichnen wir die Verteilung des Merkmals für die Kranken an der horizontalen Achse gespiegelt, damit wir in beiden Verteilungen Anteile untereinander vergleichen können. Wir sehen dadurch beide Anteile, welche einen möglichen Fehler der Diagnose widerspiegeln, weil sie sich durch den Trick nicht überlappen. Der Trennpunkt kann nach rechts oder nach links verschoben werden. Daraus ist zu erkennen, dass sich die entsprechenden Fehler gegenläufig verhalten: Wir der α-Fehler kleiner, weil der Trennbalken nach rechts geschoben wird, wird gleichzeitig der β-Fehler größer und umgekehrt. Für die Bewertung der Diagnose sind beide statistischen Fehlermöglichkeiten relevant, die falsche Diagnose „positiv“ (Überschreitung des Trennpunkts) bei Gesunden, aber auch die falsche Diagnose „negativ“ (Unterschreitung des Trennpunkts) bei Kranken. Natürlich kann man die Annahme einer Normalverteilung für das Diagnosemerkmal anzweifeln, aber für eine nicht allzu schiefe Verteilung werden sich die Schlussfolgerungen kaum wesentlich ändern.

Im medizinischen Jargon spricht man von Spezifität anstelle vom Komplement des Fehlers 1. Art, das ist der Anteil der Gesunden, die tatsächlich als Gesund erkannt werden (Diagnose negativ). Und man spricht von Sensitivität, wenn man das Komplement des Fehlers 2. Art meint, das ist der Anteil der Kranken, die tatsächlich als krank eingestuft werden (Diagnose positiv).

Güte der Diagnose im Vergleich mit Indices zur Gütebewertung von statistischen Tests

Die Güte eines potentiellen Merkmals zur Diagnose wird also, nach Fixierung eines Schwellenwerts (Trennpunkts) durch die Spezifität und die Sensitivität bewertet, also durch $1-\alpha$ und durch $1-\beta$ (auch Power, oder statistische Macht). Der Vergleich der Situation im Diagnoseverfahren und in einem statistischen Test ist aus Abb. 7 ersichtlich. Für die Bewertung der Diagnosefähigkeit eines Merkmals bewertet man allerdings den gesamten Verlauf der sogenannten ROC-Kurve, d.h., wie sich Spezifität und Sensitivität über alle möglichen Trennpunkte hinweg verhalten. Dazu findet man auch mehr in Borovcnik (2021).

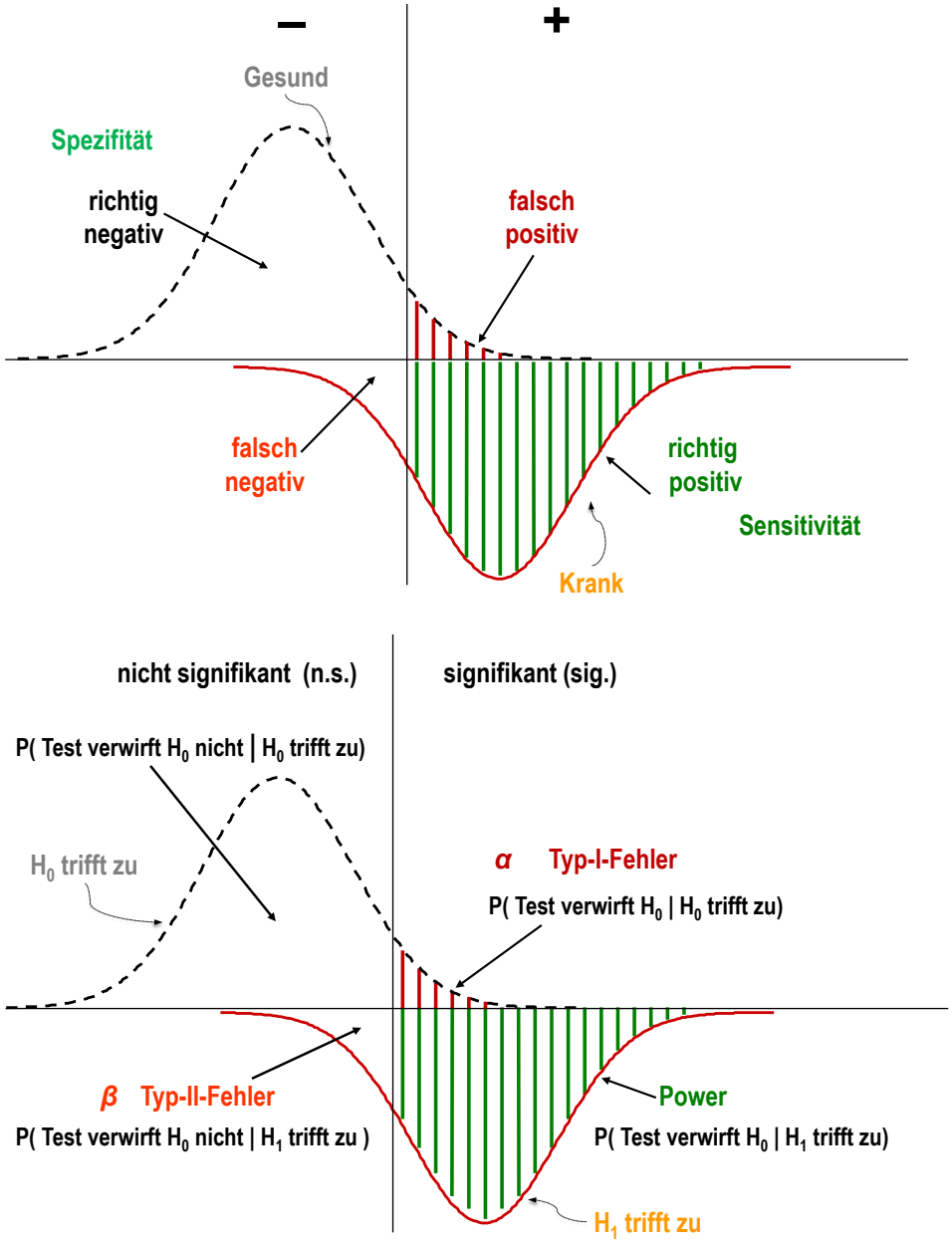

Abb. 7: Fehler bei der Diagnose und beim statistischen Test – Eine didaktisch bereichernde Analogie

2.3 Analogie zwischen der medizinischen Situation und einem statistischen Test

Fehlende Kenngrößen – wichtige Indikatoren für die Qualität von Diagnose und Tests

Wir haben bis dato die Analogie zwischen der Diagnose einer bestimmten Krankheit und einem statistischen Test so weit vorangetrieben, dass wir analoge Kennziffern zur Bewertung der Qualität des Verfahrens verwenden: Die Spezifität entspricht dem Komplement des Fehlers 1. Art, die Sensitivität dem Komplement des Fehlers 2. Art (der Macht des Tests). Allerdings hatten wir im medizinischen Kontext auch schon angesprochen, dass wir eigentlich an einer ganz anderen (bedingten) Wahrscheinlichkeit interessiert sind, nämlich an der Wahrscheinlichkeit, dass eine festgestellte Diagnose richtig ist, das ist im Falle einer positiven Diagnose der positive prädiktive Wert (*positive predictive value* PPV) sowie der negative prädiktive Wert (*negative predictive value* NPV):

$$\text{PPV} = P(\text{Krank} \mid \text{Diagnose} +)$$

$$\text{NPV} = P(\text{Gesund} \mid \text{Diagnose} -)$$

Darum geht es ja eigentlich bei der Diagnose. Und man kann diese bedingten Wahrscheinlichkeiten weder grob abschätzen, intuitiv erahnen noch kann man sie berechnen, jedenfalls kann man sie nicht berechnen, ohne dass man weitere Annahmen trifft und weitere Wahrscheinlichkeiten schätzt.

Ein Vergleich der Güte der Diagnose unter verschiedenen Rahmenbedingungen

Die Güte einer Diagnose hinsichtlich Spezifität und Sensitivität ist, egal unter welchen Rahmenbedingungen das Verfahren zur Diagnose eingesetzt wird, aus medizinischer Sicht dieselbe. Jedoch kann man daraus weder schließen, dass die Qualität dieselbe ist noch kann man die Ergebnisse miteinander vergleichen. Was soll das heißen? Das soll ein weiteres Beispiel zeigen, in dem wir Daten zur Diagnose einer bestimmten Krebsart in einer radiologischen Klinik und im Screening miteinander vergleichen. Die Daten sind in Abb. 8 enthalten, wo auch die üblichen Kenngrößen zur Bewertung der Qualität der Diagnose berechnet sind.

Aus Abb. 8 ersieht man: Die Qualitätsindices sind in beiden Situationen gleich: Sensitivität = 80%, Spezifität = 96%. Aber die Entscheidungen taugen ganz unterschiedlich: In der Klinik bietet das Ergebnis einen Anhaltspunkt: PPV = 95.2%. Im Screening ist das Ergebnis völlig obsolet und lässt keinerlei Diagnose zu: PPV 13.9%. Für die NPVs sieht die Situation kaum besser aus.

Wir können die relevanten Prädiktivwerte nur berechnen, weil wir hier die Prävalenz der Erkrankung aus den Daten direkt ablesen können. Als bedingte Wahrscheinlichkeiten interpretiert, kann man sagen, dass diese Wahrscheinlichkeiten a posteriori (nach der Diagnose, bzw. in der Analogie zum statistischen Test nach erfolgtem Test mit Ablehnung bzw. Nicht-Ablehnung) nur zu berechnen sind, wenn man die Inzidenz (Prävalenz) der Krankheit kennt (auf den statistischen Test bezogen heißt das, die a priori-Wahrscheinlichkeit der Nullhypothese).

Diese a posteriori-Wahrscheinlichkeiten haben ohne Kenntnis der a priori keinerlei Sinn und entbehren außerhalb eines engen Szenarios völlig einer Häufigkeitsinterpretation. Selbst diese Häufigkeitsinterpretation ist für den Einzelnen (Patienten) obsolet, fürs System (Gesundheitssystem) aber ein Anhaltspunkt (auch als Argumentationshilfe).

Das Missverhältnis zwischen Spezifität und Sensitivität und den Prädiktivwerten wird noch schlimmer, je kleiner die Prävalenz ist (das war bei BSE in den 2000er Jahren das Problem).

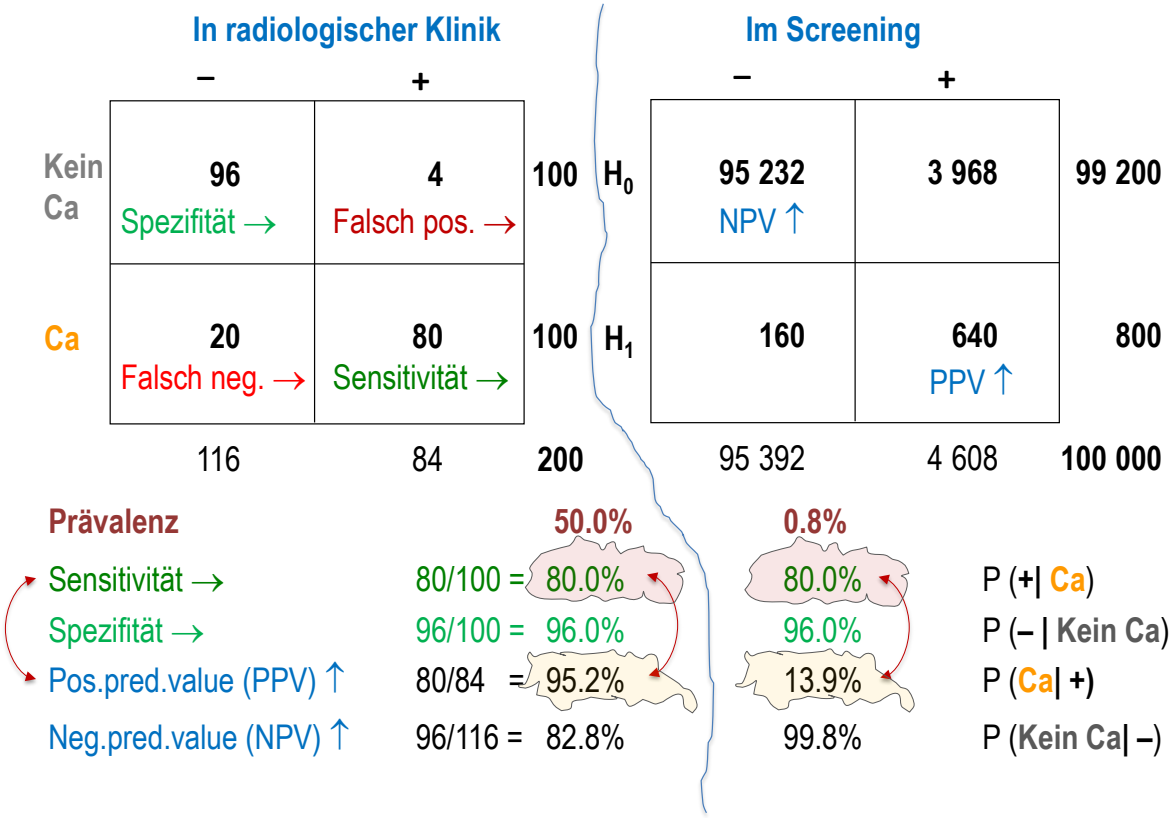

Abb. 8: Daten aus einer radiologischen Klinik und aus dem Screening zur Diagnose einer bestimmten Krebsart

2.4 Schlussfolgerungen aus der Analogie zwischen Medizin und Statistik

Im medizinischen Kontext sieht man: Der *p*-Wert ist kaum sinnvoll zu interpretieren. Die Diagnostik von Krankheiten ist ein Entscheidungsproblem, welches Verteilungen unter dem Szenario von Gesund und Krank vergleicht. Es gibt immer zwei auseinanderdriftende Fehler im Spiel:

- Diagnose der Krankheit, obwohl die Person gesund ist.
- Nicht-Erkennen der Krankheit, obwohl die Person sie tatsächlich hat.

Es gibt Krankheiten, die leicht zu diagnostizieren sind: Verschiedene Trennpunkte zum Separieren von Gesund und Krank bedingen verschiedene Größen für die Fehler. Es gibt aber neben den üblichen Modellierungsannahmen, die mehr oder weniger verletzt sind – und das ist entscheidend – einen dritten Fehler: Ob die Entscheidung gut ist, hängt nicht nur vom Trennpunkt ab, sondern wesentlich auch von der Prävalenz der Krankheit. In vielen Fällen fehlen gut interpretierbare Koeffizienten, welche die Qualität der Entscheidungen geeignet erfassen.

Die Analogie kann man mit Vorteil didaktisch verwerten: Man kann in diesem Kontext besser verstehen, was bei statistischen Tests fehlt – die a-priori-Wahrscheinlichkeit der Nullhypothese! Statistische Inferenz kann man durch die Analogie zur Medizin besser verstehen.

3. Konstituenten von Risiko-Situationen

Wir gehen auf den Charakter von Information in Zusammenhang mit Entscheidungen allgemein und in Bezug auf Gesundheitsfragen ein. Sodann unterscheiden wir Typen von Risiko, Alternativen und Evaluation und deren Auswirkung auf Entscheidungen. Wir beleuchten die Perspektiven von Entscheidungsträgern sowie den psychologischen Hintergrund, insbesondere die Logik von einzelnen und wiederholten Entscheidungen.

3.1 Information und Entscheidungen

Von einer systemanalytischen Perspektive (Borovcnik, 2011) sind auch folgende allgemeine Überlegungen zu Information als Basis für Entscheidungen entscheidend.

Was kennzeichnet Entscheidungen?

Die Aussagen, die getroffen werden

- haben was mit den Zuständen in der Wirklichkeit zu tun;
- sind auf ihren Wahrheitsgehalt zu prüfen;
- sind unvollständig oder vollständig (nur Spezialfälle? Teil von komplexen Aussagen? Etc.);
- müssen erst auf ihren Informationsgehalt interpretiert werden.
- setzen implizit Werte voraus, welche keineswegs geteilt werden müssen.

Niveaus von Information

- Daten – Fakten: – woher stammen die Daten und warum werden sie als Fakten betrachtet?
- Sind logische oder physische Zusammenhänge beteiligt?
- Daten und Resultate sind eingeschränkt auf Modelle, welche von Voraussetzungen abhängen, welche dann durch wen gesetzt werden?
- Meinungen, anekdotisches Wissen, etc.

Information: Was macht sie zuverlässig?

- Meinungsmacher („Promis“) und “Experten” – tunlichst von einer sonoren Stimme begleitet, im Extremfall ein junges Mädchen (kaum bekleidet und hilflos);
- Bilder, oder – viel besser noch – Videos;
- “Meinung der relativen Mehrheit” mit ein paar abweichenden, kritischen Stimmen;
- Wiederholung und ständiger Wechsel der Themen.

Exemplarische Beispiele: Neue Grippe, BSE, Vogelgrippe, HPV, Klimawechsel, Covid 19

- Spärliche Daten; partielle Information, losgelöst von ihrem Umfeld;
- Meinungsmacher und „Experten”;
- Bilder, von Opfern, etc.;
- Wiederholung und Wechsel.

3.2 Information in Gesundheitsfragen:

Wie Information eingefordert und evaluiert wird, wie sie in Entscheidungen einfließt, hängt vom Typ des Risikos und von im Modell berücksichtigten Alternativen ab (siehe auch Borovcnik, 2019b).

Typen von Risiko

Direkte Risiken: Persönlicher Zustand ist schwierig; Symptome oder Einschränkungen sind da, etc. Erwartung: es muss rasch helfen, es darf keine Anstrengungen erfordern. Evaluation der Maßnahmen: kurz- oder langfristig; Symptome verschwinden kurzfristig – aber, kommen sie zurück?

Virtuelle Risiken: Persönlicher Zustand ist in Ordnung. Es stehen keine Entscheidungen an, aber es lauern Gefahren. Szenarien werden „gezeichnet“: Absichern oder Panikmache? Eine Evaluation ist nur hypothetisch, weil sich die Gefahren auf die Zukunft beziehen.

Gesellschaftliche Risiken: Weder persönlichen Risiken noch Entscheidungen. Gefahr ist virtuell, untermauert durch Szenarios. Das Risiko hat jedoch eine großflächige Auswirkung: BSE, Kernenergie, Klimawechsel, etc. Eine Evaluation ist aber unmöglich.

Entscheidung: Raum der Alternativen und Evaluation

Reale Probleme: Welche 'Behandlungen' kommen in Frage? Welche komplementären 'Behandlungen' sind vielversprechend?
Virtuelle zukünftige Risiken: Welches Szenario erweist sich als wahr? Welche Maßnahmen oder Vorkehrungen sind möglich?
Gesellschaftliche Risiken: Was sind die Alternativen? Sozialer Druck – Verantwortlichkeit – Vorteile für einzelne Stakeholder?

Gute Information, um die Entscheidung zu stützen, ist schwer zu bekommen und schwer zu verstehen; sie bezieht sich auf Mathematik und Statistik: virtuelle Aussagen in Szenarios; sie ist schwierig zu evaluieren und erfordert 'Expertise'; sie spiegelt Interessen jener, welche die Information liefern. Kann man auf den Rat von Experten vertrauen? Sind das Experten für kurz- oder langfristige Entwicklungen? Haben die Experten Interessenskonflikte?

3.3 Perspektiven von Entscheidungsträgern

Das Zeitalter der Versicherung und Rückversicherung

Wir leben im Zeitalter der Versicherung – wir versichern Auto, Leben, Haus, Ernte, etc. In einer mechanistischen Sicht der Welt werfen wir eine Münze in den Automaten und erwarten, genau das zu bekommen, was wir benötigen, und das ohne jede weitere Anstrengung: Wir haben Kopfweh – wir nehmen eine Tablette Paracetamol. Wir leiden an depressiver Verstimmung: wir nehmen ein Prozac. Wir sind im Sport hintennach: wir nehmen Cera. Etc. Wir benehmen uns wie Bürger in Orwells 1984 "Big Brother's watching you". Aber wir sind überrascht, wenn man uns permanent kontrolliert.

Die Medizin steht unter enormem Leistungsdruck

Wer nicht alles verspricht, wird Kunden verlieren. Die anderen „draußen" bieten alles an. Je mehr wir uns dem Lebensende nähern, umso härter wird dieser Wettbewerb. Aber: Ist Wettbewerb immer schlecht? Komplementäre Maßnahmen könnten die Wirkung von medizinischer Behandlung richtig boostern. Wie soll man ein Versprechen zu helfen evaluieren? Fragen, die auf jeder Checkliste stehen sollten: Wer bietet ausreichende Information, um unabhängig zu entscheiden? Wer versteht die angebotene Information? Was charakterisiert Information? Nach welchen Kriterien wird sie beurteilt?

Stakeholder und ihre divergierenden Interessen

Medizin: Behandlung lege artis muss garantiert werden. – Ärzte sind haftbar und werden jegliches „Risiko" vermeiden.
Experten: Benötigen Forschungsgelder – haben ihre Rolle als wichtig darzustellen. – Sie tragen keinerlei Risiken.
Gesundheitssektor: Zwischen Pharma und öffentlicher Meinung. – Trägt keine Risiken, aber erfährt eine Art von „Verantwortung".
Medien: Sind – mit Ausnahmen – eher an Sensationen interessiert. Sie tragen keinerlei Risiken.
Politiker: Müssen sich als Macher erweisen, die rasch gute Entscheidungen treffen. – Sie tragen das Risiko ihrer persönlichen Karriere.
Selbsthilfegruppen: Haben verschiedenste Interessen und müssen sich wichtigmachen. – Sie tragen keinerlei Risiken.
Pharma-Industrie: Hat ein natürliches Interesse an Geschäft und Profit. – Ist haftbar. Muss Verfahren vor Gericht gewärtigen (Contergan).

Mehr zu den Stakeholdern und ihren Interessen findet man in Borovcnik und Kapadia (2011). Der Impakt in Form des Schadens, das eigentliche Risiko, und das muss man im Auge behalten, verbleibt immer beim Individuum.

3.4 Der psychologische Hintergrund

Auffassungsunterschiede, Nutzen und Wahrscheinlichkeit, das Feld der Widersprüche ist weit. Die Situation wird schwierig und das Verhalten mag beeinflusst werden, wenn Geld im Spiel ist; wenn die Beträge viel größer sind; wenn jemand die Ausgangssituation (zu Recht?) anders sieht.

Ein klassisches Beispiel sind die Experimente von Kahneman und Tversky (1979) mit Erwachsenen, die letztlich in die Interpretation des Verhaltens von Probanden als Risikoscheu in Gewinnsituationen und Risikosuchend in Verlustsituationen gemündet hat, was in der *Prospect theory* verarbeitet wurde und letztlich auch zur Nobelpreisverleihung von Kahneman in Wirtschaftswissenschaften geführt hat. Allerdings hat Borovcnik (2015, 2016a) die Experimente re-analysiert und kommt zu einer gegenläufigen Einschätzung, wonach die Menschen zu Recht das Risiko vermeiden, wenn sie eigentlich einen größeren Bestand ihres Vermögens aufs Spiel setzen würden, wogegen sie das Risiko suchen, wenn sich eine Gelegenheit ergibt, einen bestehenden Stand von Schulden mit einem Mal abzubauen.

Man sieht also, dass nicht nur der Nutzen den Blickwinkel verändert, sondern man muss zugestehen, dass Menschen zu Recht eine Situation verschieden auffassen können und sich selbst Nobelpreisträger irren können. Dieses Verhalten spiegelt sich auch darin, dass Menschen in der Lotterie und im Lotto mitspielen und dass sie jegliches Risiko durch Versicherungen abzudecken versuchen. Wobei sie übersehen, dass die Versicherung einen Vertrag ja nicht ohne Gewinnabsicht anbieten wird, ja nicht anbieten kann.

Das am schwersten zu akzeptierende ist allerdings, dass eine für die *wiederholte* Entscheidungssituation abgestimmte Entscheidung zu einer anderen Entscheidung führt, als wenn man nur *eine* Entscheidung treffen muss – ja die Entscheidungen kehren sich geradewegs um. Darüber hinaus ist festzustellen, dass bei wiederholten Entscheidungen der Spielraum des Zufalls völlig zusammenbricht, während man bei einer Einzelentscheidung das volle Risiko der Zufallsschwankungen zu gewärtigen hat. Da spiegelt sich auch darin, dass ein einzelner Versicherungsnehmer erst am Ende der Versicherungsperiode weiß, ob seine Entscheidung zur Versicherung oder dagegen richtig war, die Versicherung aber Verträge anbieten kann und sehr genau das wirtschaftliche Ergebnis ohne weitere Risiken abdecken kann. Wenn einzelne Versicherungsgeschäfte über das übliche Maß hinausgehen, so schließen sich die Versicherungen über Rückversicherer zusammen, um wieder die Situation einer wiederholten Entscheidung mit Ausgleich des Risikos herzustellen.

4. Schlussfolgerungen für den Umgang mit Risiken

Informationen über Risiken – Bedarf an innovativen Wegen

Wesentliche Merkmale von Risiken sind: Risiken werden mit Wahrscheinlichkeiten ausgedrückt; mit Risiken sind oft sehr kleine Wahrscheinlichkeiten verbunden. Es mangelt an Daten zum Risiko; es ist oft schwierig, die erforderlichen Daten zu erhalten, nicht nur für kleine Wahrscheinlichkeiten. Die Ergebnisse von Analysen sind komplex, basieren auf Modellen und haben einen gewissen Interpretationsspielraum, sodass selbst Experten sich vielmals in Schwierigkeiten finden.

Patienten wünschen sich mehr Informationen über Risiken. Aber: Verstehen Patienten und Ärzte wirklich die Implikationen? Wie kann Information zwischen den ungleichen Stakeholdern verantwortlich geteilt werden? Es bedarf neuer Wege der Informationsvermittlung. Da gibt es noch viel zu verbessern. Das Gesundheits“system” reagiert unterschiedlich: Es verweigert innovative Lösungen; oder es „informiert“ die Patienten viel besser, bürdet ihnen aber eine unzumutbare Verantwortung auf.

Risikokompetenz: Der informierte Patient

Was die Kommunikation zwischen Patienten und Ärzten anbelangt: Ist sie unzureichend, erhöht sich das Risiko von Fehlbehandlungen. Bessere Information kann aber das Vertrauen zwischen Patient und Arzt und zwischen den beteiligten Stakeholdern auch beeinträchtigen (und einen Placebo-Effekt verhindern). Ferner: Nicht alle Patienten wollen selbst entscheiden (das erfordert neben der mühseligen Beschäftigung auch die Übernahme von Verantwortung). Viele Patienten wünschen sich eine gemeinsame Entscheidung. In der Realität muss man aber kritisch hinterfragen: Kann der Patient die relevante Information wirklich bewerten? Kann der Arzt die relevante Information ordentlich vermitteln?

Wie können Risiken in Gesundheitsfragen besser kommuniziert werden? Es gibt mehrere Studien, zu innovativen Formaten zur Verbesserung der Risikokommunikation (siehe die Ausführungen in Abschnitt 1, auch Gigerenzer, 2002, 2014, oder Harding Center for Risk Literacy, o.D.).

Risikokompetenz in Gesundheitsfragen – die Experten herausfordern

Qualität und Quantität von Informationen: Mehr „Information" bedeutet oft weniger Information. Es geht darum, neue Wege zu finden, um Experten herauszufordern, ihnen die entscheidenden Fragen zu stellen, und ihre Antworten zu beurteilen, ohne sich ihre Expertise anzueignen, was völlig außer Reichweite wäre. Die Unterstützungssysteme für Patienten könnten durchaus ausgebaut werden. Wenn es der Wunsch des Patienten ist, könnte das Ziel sein, für eine gemeinsame Entscheidung (shared decision) mitverantwortlich zu sein, auch wenn die Interessen und die Folgen zwischen den an diesem Entscheidungsprozess beteiligten Stakeholdern völlig konträr sind.

Abschließend fünf Thesen zum Risiko, die auch den Umgang mit Risiko in der Medizin bestimmen

Wahrscheinlichkeit ist ein virtuelles Konzept. Insbesondere kleine Wahrscheinlichkeiten entziehen sich jeder frequentistischen Interpretation.

Es gibt keine Rationalität ohne Vergleich. Um die Wahrscheinlichkeit für eine Entscheidungsfindung nutzbar zu machen, braucht man mehrere Optionen, zwischen denen man wählen kann. Man muss Risiken vergleichen, anstatt Risiken zu berechnen.

Die Optimalität von Entscheidungen ist nicht klar definiert. Die Optimalität hängt vom Stakeholder, den verwendeten Zielkriterien, der Interpretation von Wahrscheinlichkeit und von Wertevorstellungen, auf die man sich stützt, ab.

Informationen sind für die beteiligten Akteure unterschiedlich nützlich. Individuen sind ganz anders betroffen als das System.

Einmalige und wiederholte Entscheidungen folgen einer unterschiedlichen Logik. Was bei wiederholten Entscheidungen eine gute Option sein mag, kann im Einzelfall sehr schlecht sein. Die unterschiedliche Logik macht einen Versicherungsvertrag praktikabel, lässt sich aber nicht auf Gesundheitsentscheidungen zwischen System und Individuum übertragen.

Ich danke Franz Pauer für seine kritischen Anmerkungen, welche die Verständlichkeit meiner Ausführungen sehr verbessert hat.

Literatur

ASA (2016): Statement on statistical significance and p-values. *The American Statistician* 70(2), 131–133. amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.XKKMv9hS-Hs

Batanero, C. & Borovcnik, M. (2016): *Statistics and probability in high school*. Rotterdam: Sense Publishers.

Borovcnik, M. (2015): Risk and decision making: The "logic" of probability. *The Mathematics Enthusiast 12*(1–3), 113–139.

Borovcnik, M. (2016a): Risiko – ein Überlebensratgeber. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 49*, 1–16.

Borovcnik, M. (2016b): "To screen or not to screen" … Dialoge zur medizinischen Diagnose. mathematik lehren 194, 22–28.

Borovcnik, M. (2019a): Risk and Decision Making – Modeling and Statistics in Medicine: Case Studies. In B. Sriraman (Hrsg.): *Handbook of the Mathematics of the Arts and Sciences*. Springer Nature. doi.org/10.1007/978-3-319-70658-0_62-1

Borovcnik, M. (2019b): Risk and Decision Making – Modeling and Statistics in Medicine: Fundamental Aspects. In B. Sriraman (Hrsg.): *Handbook of the Mathematics of the Arts and Sciences*. Springer Nature. doi.org/10.1007/978-3-319-70658-0_62-1

Borovcnik, M. (2021): Informelle statistische Inferenz. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 53*, 19–34.

Borovcnik, M. & Kapadia, R. (2011): Risk in health: more information and more uncertainty. In: *Proceedings IASE Satellite Conference on "Statistics Education and Outreach"*. Voorburg: ISI. iase-web.org/Conference_Proceedings.php?p=Stats_Education_and_Outreach_2011

Borovcnik, M. & Kapadia, R. (2018): Reasoning with risk: Teaching probability and risk as twin concepts. In: Batanero, C., Chernoff, E. J., Engel, J., Lee, H. & Sánchez, E. (Hrsg.): Research on teaching and learning probability.New York: Springer, 3–22.

Dubben, H.-H. & Beck-Bornholdt, H.-P. (2010): Der Hund, der Eier legt. Erkennen von Fehlinformation durch Querdenken. Reinbek: Rowohlt.

Gigerenzer, G. (2002): *Calculated risks: How to know when numbers deceive you*. Simon & Schuster.

Gøtzsche P. C. & Nielsen M. (2011): Screening for breast cancer with mammography – Review. *The Cochrane Library 4*. doi.org/10.1002/14651858.CD001877.pub4

Gigerenzer, G. (2014): Risiko. Wie man die richtigen Entscheidungen trifft. München: Random House. Original: Risk savvy. How to make good decisions. New York: Penguin.

Harding Center for Risk Literacy (o.D.). *Early detection of breast cancer by mammography screening*. www.hardingcenter.de/en/early-detection-of-cancer/early-detection-of-breast-cancer-by-mammography-screening

Hubbard R. & Bayarri M. J. (2003): Confusion over measures of evidence (p) versus errors (α) in classical statistical testing. *The American Statistician 57*(3), 171–182.

Kahneman, D. & Tversky, A. (1979): Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Koubenec, H. J. (2000): Mammographie-Screening: Überschätzen wir den Nutzen? (Mammography screening: Do we overestimate the benefits?) *Berliner Ärzte 37*(8), 11–16.

Mongin, P. (1997): Expected utility theory. In: Davis J, Hands W, Maki U (Hrsg.): *Handbook of economic methodology*. London: Edward Elgar, 342–350.

Sandblom, G., Varenhorst, E., Rosell, J., Löfman, O. & Carlsson, P. (2011): Randomised prostate cancer screening trial: 20year follow-up. *British Medical Journal Online 342*. doi.org/10.1136/bmj.d1539

Spiegelhalter, D. (2012): Using speed of ageing and "microlives" to communicate the effects of lifetime habits and environment. *British Medical Journal* 345:e8223. doi.org/10.1136/bmj.e8223

Spiegelhalter D. (2014, April): What can education learn from real-world communication of risk and uncertainty? Invited lecture at the Eight British Congress on Mathematical Education, Nottingham.

Spiegelhalter, D. & Gage, J. (2015): What can education learn from real-world communication of risk and uncertainty? *The Mathematics Enthusiast* 12(1-3), 4–10. scholarworks.umt.edu/tme/vol12/iss1/4/

Wasserstein R. L. & Lazar N. A. (2016): The ASA's statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–131. amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.XKKMv9hS-Hs

Wewetzer, H. (2005): „Der Nutzen ist fraglich". Der Psychologe Gerd Gigerenzer über Sinn und Unsinn der Brustkrebs-Reihenuntersuchungen. Der Tagesspiegel, 1.6.2005. www.tagesspiegel.de/themen/gesundheit/der-nutzen-ist-fraglich/612902.html

Verfasser

Manfred Borovcnik
Universität Klagenfurt, Institut für Statistik, Sterneckstraße 15, 9020 Klagenfurt
manfred.borovcnik@aau.at