Das Fehlalarm-Paradoxon
Warum produziert ein guter Test so miserable Ergebnisse?

Klaus Pommerening

Mai 2020, letzte Änderung: 12. November 2020

Das Problem[1]

Nehmen wir an, für eine Infektion, nennen wir sie CoV-2, wird ein Labortest an vielen Menschen durchgeführt. Dieser Test hat hervorragende Qualitätskennzahlen:

- Er erkennt bei 98 % der Infizierten die Infektion korrekt. Der Fachausdruck dafür ist **Sensitivität** – diese ist also[2] 98 %.
- Er schlägt nur bei 5 % von Nichtinfizierten an, erkennt also die Nichtinfizierten zu 95 % richtig. Der Fachausdruck dafür ist **Spezifität** – diese ist also 95 %.

Nun wird eine Person getestet, und der Test fällt positiv aus. *Wie groß ist die Wahrscheinlichkeit, dass diese Person infiziert ist?* 95 % oder 98 %?

Weit gefehlt – keine der beiden Antworten kommt auch nur in die Nähe der richtigen Lösung!

Ein Zahlenbeispiel

In Wirklichkeit kann die Frage ohne eine weitere Information gar nicht beantwortet werden, eine Information, die mit der Qualität des Tests *nichts zu tun hat*, für seine Ergebnisse trotzdem entscheidend ist: Wieviele Personen, egal ob schon getestet oder nicht, sind insgesamt infiziert?

Nehmen wir an, von der Gesamtbevölkerung von 80 Millionen eines Staates seien 160 000 aktuell infiziert[3]. Der Fachausdruck dafür ist **Prävalenz** – diese ist in diesem Zahlenbeispiel also 160 000 geteilt durch 80 Millionen, macht 0,002, das sind 2 ‰.

Welche Ergebnisse können wir dann bei einem Test von 100 000 Personen erwarten? Von diesen sind wahrscheinlich etwa 200 infiziert. Bei diesen ist der Test wahrscheinlich in 196 Fällen positiv, in vier Fällen negativ.

[1] Eine Zusammenfassung der Kernaussagen für die aktuelle Corona-Pandemie findet sich am Ende dieses Artikels.

[2] Man ist natürlich versucht zu fragen: Warum nicht 100 %? Nun, erstens können Fehler auftreten, angefangen von der Probenentnahme bis hin zur Fehleinschätzung einer Färbung im Reagenzglas. Außerdem nimmt man bei höherer Sensitivität in der Regel in Kauf, dass mehr Nichtinfizierte fälschlicherweise ein positives Testergebnis erhalten, dass also die Spezifität schlechter ist. Ideal wäre natürlich ein Test, der Infizierte und Nichtinfizierte sicher auseinanderhalten kann, also jeweils 100 % Sensitivität und Spezifität hat – aber das ist in der Realität illusorisch.

[3] Mitte November 2020 gibt es in Deutschland rund 700 000 bisher positiv auf SARS-CoV Getestete. Das dürfte, siehe weiter unten, auf eine Zahl von 100 000 bis 200 000 Infizierten hindeuten. Das Beispiel wird aber auch später in diesem Artikel mit einer Infiziertenzahl von 700 000 durchgerechnet, siehe den Abschnitt „Eine Gegenrechnung“.

Die übrigen 99 800 unter den Testpersonen sind von der Infektion verschont. Diese bekommen zu 95 % das korrekte negative Testergebnis, bei den übrigen 5 % ist das Ergebnis positiv, also bei 4 990, gegenüber der überwältigenden Mehrheit von 94 810 korrekt negativen Ergebnissen. Abbildung 1 veranschaulicht die Situation.

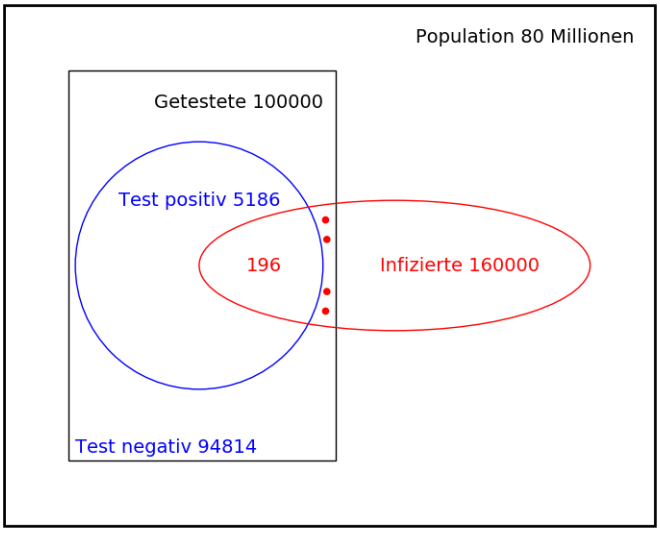

Abbildung 1: Testergebnisse (fiktiv) – die Größe der Flächen ist nicht proportional zu den jeweils repräsentierten Anzahlen; der konkavlinsenförmige Abschnitt besteht aus den 4 Infizierten (rote Punkte), die beim Test als negativ durchrutschen.

Um die Auswirkung dieser Überlegungen deutlich zu erkennen, werden die in etwa zu erwartenden Ergebnisse in eine **Vierfeldertafel** eingetragen:

CoV2-Test	infiziert? ja	nein	Summe
+	196	4 990	5 186
–	4	94 810	94 814
Summe	200	99 800	100 000

Wir haben also 5 186 positive Testergebnisse zu erwarten, darunter aber nur 196 korrekte bei 4 990 (!) falschen. Das bedeutet (in diesem Zahlenbeispiel):

> *Die Wahrscheinlichkeit, dass eine positiv getestete Person tatsächlich infiziert ist, ist 196/5186, also ungefähr 3,8%.*

Diese Größe nennt man den **Vorhersagewert** oder genauer: den positiven Vorhersagewert.

> *Die ungefähr 96,2% positiven Ergebnisse des Tests bei Nichtinfizierten bedeuten hingegen jeweils einen* **Fehlalarm.**

Und dass der Vorhersagewert eines positiven Testergebnisses so verblüffend gering ist, ist das **Fehlalarm-Paradoxon**.

Erklärung

Woran liegt dieses „Versagen“? Nun, die vielen, nämlich fast 5 000, „falsch positiven“ Testergebnisse kamen einfach dadurch zustande, dass die Anzahl der Nichtinfizierten so viel größer ist als die Zahl der Infizierten und damit der geringe Prozentsatz an falsch positiven Ergebnissen trotzdem in absoluten Zahlen beträchtlich ist – jedenfalls viel größer als die Zahl der mit hoher Wahrscheinlichkeit gefundenen „echt positiven“ Ergebnisse.

Der Schluss von der Ursache auf die Wirkung lässt sich eben nicht einfach umkehren zu einem Schluss von der Wirkung auf die Ursache! Mehr dazu – mit etwas mathematischer Untermauerung – in einem späteren Abschnitt.

Sehen wir uns als Gegenprobe ein Zahlenbeispiel an, bei dem ungefähr die Hälfte der untersuchten Bevölkerung infiziert ist. Dann sähe die Vierfeldertafel (bei gleicher Sensitivität und Spezifität des Tests) etwa so aus:

	infiziert?		
Test	ja	nein	Summe
+	49 000	2 500	51 500
–	1 000	47 500	48 500
Summe	50 000	50 000	100 000

Von den 51 500 postiv getesteten wären also 49 000 tatsächlich infiziert – der positive Vorhersagewert läge bei 49 000/51 500, also bei 95 %. Die Fehlalarmquote in diesem Beispiel wäre 2 500/51 500, also etwa 5 %. Das sähe akzeptabel aus.

Der Haken bei der Sache ist nur, dass (glücklicherweise) eine so hohe Zahl von tatsächlich Infizierten in der Realität so gut wie nie vorkommt.

Wichtig zu wissen ist jedenfalls, dass der miserable Vorhersagewert nicht ein Qualitätsmanko des Tests ist, sondern *durch die geringe Prävalenz der nachzuweisenden Infektion bedingt ist*. Unter einer solchen Fehleinschätzung leidet immer wieder das Ansehen vieler wichtiger Tests und Diagnoseverfahren, z. B. des Mammographie-Screenings.

Auswege aus der Falle?

Einen einfachen Ausweg gibt es nicht, die Zahlen sprechen für sich. Je seltener eine Infektion oder Erkrankung in der Bevölkerung ist, desto mehr Fehlalarme, also falsch positive Testergebnisse, muss man in Kauf nehmen. Man könnte eventuell diese Zahl senken, indem man die Spezifität des Tests erhöht. Man muss dann aber in Kauf nehmen, dass er mehr falsch negative Ergebnisse produziert, dass einem also Infizierte „durch die Lappen“ gehen. Das dürfte in den meisten Fällen keine reale Option sein, denn das hauptsächliche Ziel ist ja gerade, die Infizierten zu finden.

Nehmen wir trotzdem an, die Spezifität könnte erhöht werden. Das könnte etwa geschehen, indem man einen Grenzwert in Richtung „weniger positive Ergebnisse“ verschiebt – z. B., wenn man den Grad der Färbung in einem Reagenzglas, der einen positiven Befund anzeigen soll,

etwas höher ansetzt. Könnte man bei unserem Zahlenbeispiel eine Spezifität von 98 % bei einer Sensitivität von 95 % erreichen, so sähe die Ergebnistafel so aus:

CoV2-	infiziert?		
Test	ja	nein	Summe
+	190	1 996	2 186
–	10	97 804	97 814
Summe	200	99 800	100 000

Die Fehlalarmrate dieses modifizierten Tests wäre dann 1 996/2 186, also immer noch über 91 %, und der Vorhersagewert wäre auf knapp 9 % erhöht. Der Preis dafür wäre allerdings, dass 10 statt nur 4 wirklich Infizierte übersehen werden.

Was bleibt an möglichen Auswegen?

- Natürlich hilft es, wenn ein besserer Test angewendet werden kann, also mit erhöhter Spezifität, und dabei mindestens gleich guter Sensitivität. Das ist aber höchstens dann eine realistische Option, wenn es diesen besseren Test schon gibt, er aber wegen seiner Verfügbarkeit, seiner Belastung für die zu testenden Personen oder seiner Kosten zunächst nicht eingesetzt wird. Und wie gesehen ist der Vorhersagewert trotzdem nicht wirklich viel besser, weil er auch für den besseren Test sehr stark von der „Durchseuchung“ der Bevölkerung abhängt.

- Das wichtigste ist auf jeden Fall, dass die Ärztin, die dem Betroffenen das Testergebnis mitteilt, sich der Problematik bewusst ist und den Vorhersagewert in etwa kennt. Sie sollte dem Patienten also nicht sagen: „Sie haben Krebs.“ – sondern: „Das Testergebnis ist positiv, aber die Wahrscheinlichkeit, dass Sie wirklich Krebs haben, liegt bei etwa 4 %. Wir sollten das natürlich weiter überprüfen.“

- Deutlich höher ist der Vorhersagewert, wenn aus anderem Zusammenhang schon ein Verdacht auf die Infektion besteht, wenn der Test also nur bei einem Teil der Bevölkerung durchgeführt wird, bei dem die Infektion wesentlich häufiger auftritt. Bei einer Infektion, die durch die Atemluft auf Personen in geringem Abstand übertragen wird, würden beispielsweise nur Fälle getestet, die direkten Kontakt mit einem bereits Infizierten hatten.

Fällen wir für diese Situation einer „vorgesiebten“ Testpopulation wieder eine Vierfeldertafel aus. Dabei wird angenommen, dass das etwa 10 000 Personen sind und unter diesen etwa 200, also 2 % statt nur 2 ‰ infiziert sind[4].

	infiziert?		
Test	ja	nein	Summe
+	196	490	686
–	4	9 310	9 314
Summe	200	9 800	10 000

[4]Die Vorauswahl hat in diesem Zahlenbeispiel also alle Infizierten erfasst. Das entspricht einem „Vortest“ mit 100 % Sensitivität, aber einer sehr niedrigen Spezifität von nur 2 %.

Von den 686 positiv getesteten sind 196 wirklich infiziert, d. h., der Vorhersagewert ist jetzt 196/686, liegt also bei knapp 29 %. *Das ist immer noch sehr bescheiden, aber doch deutlich besser als der ursprüngliche Wert von 3,8 %.*

An der Zahl der 4 nicht entdeckten Infizierten hat sich übrigens nichts geändert – diese kommt ja durch die Sensitivität von 98 % zustande.

Die Lehre aus dieser Analyse mit Zahlenbeispiel ist also:

> *Ein positives Testergebnis ist meistens noch weit von einer endgültigen Diagnose entfernt. Es muss so gut wie immer durch weitere Untersuchungen überprüft werden, um zu einer gesicherten Diagnose zu kommen.*

Hier noch eine ergänzende Bemerkung: Wie sicher kann sich eine *negativ* Getestete sein, dass sie nicht infiziert ist? Dies wird durch den **negativen Vorhersagewert** ausgedrückt: In der Vierfeldertafel im ersten Zahlenbeispiel entspricht das dem Wert 94 810/94 814, der erfreulicherweise praktisch 100 % bedeutet.

Was bringt ein zweiter Test?

Das Szenario „Erst ein Vortest, dann ein weiterer Test“ passt besonders für den Fall, dass es zwei unabhängige Tests für die Infektion gibt. Dann kann das Ergebnis von Test 1 als Vorauswahl für Test 2 dienen. Die positiv Getesteten von Test 1 bilden in diesem Szenario die zu untersuchende Bevölkerung von Test 2, und in dieser Gruppe ist die Prävalenz wesentlich höher – im obigen Zahlenbeispiel 196 von 5 186, also[5] 3,8 %. Das ist entscheidend mehr als die Prävalenz von 2 ‰ in der Gesamtbevölkerung.

Füllen wir für die Situation eines zweiten, unabhängigen Tests auch wieder eine Vierfeldertafel aus, für die 5 186 Testpersonen, die beim Test 1 auffällig waren. Ansonsten gehen wir für das Beispiel als Qualitätskenngrößen des zweiten Tests von 95 % Sensitivität und 98 % Spezifität aus, bei einer Prävalenz von 3,8 % in der untersuchten Gruppe:

	infiziert?		
Test	ja	nein	Summe
+	186	100	286
–	10	4890	4 900
Summe	196	4 990	5 186

Von den 286 positiv getesteten sind 186 wirklich infiziert, d. h., der Vorhersagewert ist jetzt 186/286, liegt also bei 65 %. Das ist wesentlich besser, hat aber einen Preis: Weitere 10 Infizierte sind als nichtinfiziert klassifiziert worden. *Das ist in der Regel nicht hinnehmbar. Wenn ein zweiter Test oder eine weitere Untersuchung angeschlossen wird, sollte diese eine Sensitivität von 100 % haben!* Jede Abweichung davon lässt weitere Infizierte unentdeckt. Der zweite Test muss also *sehr* exakt sein und sehr sorgfältig durchgeführt werden, was in der Regel mit erheblichem Aufwand verbunden ist.

Der *negative* Vorhersagewert des zweiten Tests liegt übrigens immer noch bei knapp 100 %. Das bedeutet: Wer beim ersten Mal positiv, beim zweiten aber negativ gestestet wird, ist mit ziemlicher Sicherheit nicht infiziert.

[5] Es ist kein Zufall, dass das genau mit dem Vorhersagewert von Test 1 übereinstimmt.

Eine Gegenrechnung

Die Annahmen im Zahlenbeispiel waren eher „Corona-skeptisch“, indem sie die Prävalenz und die Genauigkeit des Tests niedrig ansetzten. Leider gibt es dafür keine genauen, belastbaren Werte, was bedeutet, dass die daraus gezogenen Folgerungen auch nicht exakt sein können.

In einer solchen Situation – einer mathematischen Modellrechnung mit bestenfalls grob exakten Werten – hilft man sich, indem man die Annahmen über die verwendeten Parameter variiert. Da die erste Rechnung mit „Corona-skeptischen“ Parametern durchgeführt wurde, folgt jetzt eine Rechnung, die das entgegengesetzte Extrem nachbildet:

- Die Zahl der tatsächlich Infizierten wird als 700 000 angenommen. Das bedeutet, dass sich die falsch positiven Ergebnisse und die Dunkelziffer der nicht getesteten Infizierten in etwa ausgleichen. Die Prävalenz liegt unter dieser Annahme also bei 700 000/80 000 000 $\approx 9\,‰$.
- Die Sensitivität wird mit 98 % angenommen,
- die Spezifität mit 99 %.

Exakte Angaben zu Sensitivität und Spezifität scheinen bisher nicht zu existieren. In einem Artikel im Deutschen Ärzteblatt[6] wird für die Sensitivität eine breite Spanne mit maximal 98 % konstatiert. Dass dieser Maximalwert wohl noch zu hoch gegriffen ist, liegt daran, dass Fehler bei der Prozessierung, z. B. beim Abstrich, hier voll durchschlagen. Die Spezifität liegt nach Angabe einer Laborleiterin[7] bei 99%. Ansonsten findet man hauptsächlich Aussagen der Art: „Die Genauigkeit des PCR-Tests liegt bei nahezu 100%“ – eine sehr ungenaue Angabe, bei der nicht einmal zwischen Sensitivität und Spezifität unterschieden wird und die daher unbrauchbar ist.

Nun, unter diesen, je nach Gesichtspunkt sehr optimistischen oder sehr pessimistischen, Annahmen sieht die Vierfeldertafel so aus:

	infiziert?		
Test	ja	nein	Summe
+	882	991	1 873
−	18	98 109	98 109
Summe	900	99 100	99 100

Daraus errechnet sich die Fehlalarmrate zu $991/1873 \approx 0.52$ oder 52%. D. h., *selbst unter diesen extremen Annahmen ist immer noch mehr als jedes zweite positive Testergebnis ein Fehlalarm.*

---

[6] Ralf L. Schlenger: PCR-Tests auf SARS-CoV-2: Ergebnisse richtig interpretieren. Dtsch Arztebl 2020; 117(24): A-1194 / B-1010 – Artikel im Deutschen Ärzteblatt werden vor ihrer Publikation einem Review-Prozess nach wissenschaftlichen Standards unterworfen.

[7] Südkurier vom 11. November 2020 – Dass sie daraus einen Vorhersagewert von ebenfalls 99% folgert, spricht allerdings für mangelnde statistische Kenntnisse.

Im nächsten Abschnitt werden diese an Zahlenbeispielen gewonnenen Erkenntnisse mit etwas mehr Mathematik unterfüttert.

Etwas Theorie: Die Formel von Bayes

Das Problem des Umkehrschlusses von der Wirkung auf die Ursache wird in der Wahrscheinlichkeitsrechnung durch die Formel von Bayes beschrieben. Zum Verständnis ist Vertrautheit mit dem (aus der Schulmathematik bekannten) „algebraischen“ Ansatz, unbestimmte Zahlenwerte durch Buchstaben wiederzugeben, nötig. Außerdem wird der mathematische Begriff der Wahrscheinlichkeit verwendet, wobei hier die naive Vorstellung von Wahrscheinlichkeit als relativer Häufigkeit ausreicht[8].

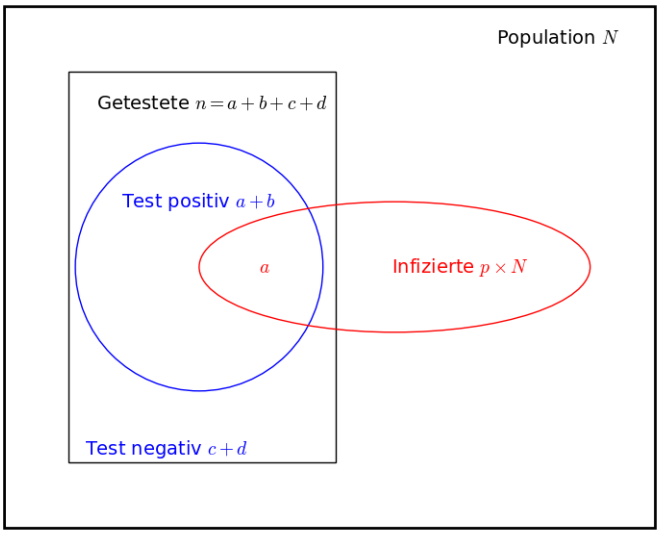

Abbildung 2: Testergebnisse abstrakt – der konkavlinsenförmige Abschnitt besteht aus den c Infizierten, die beim Test als negativ durchrutschen.

Die Überlegung, veranschaulicht durch Abbildung 2, startet wieder mit der Vierfeldertafel der Testergebnisse, diesmal mit unbestimmten Werten ausgefüllt:

[8]Wir stellen uns eine Menge aus n Elementen als Sammlung von Kugeln in einer Urne vor, rote und weiße. Die Teilmenge $B \subseteq A$ soll genau aus den m roten Kugeln bestehen. Dann ist die Wahrscheinlichkeit $P(B)$, bei zufälliger Auswahl einer Kugel eine rote zu erwischen, gerade die relative Häufigkeit m/n.

Test	infiziert? ja	nein	Summe
+	a	b	$a+b$
−	c	d	$c+d$
Summe	$a+c$	$b+d$	n

Dabei ist $n = a + b + c + d$ die Gesamtzahl der untersuchten Population, also der Getesteten. Die relevanten Größen werden jetzt durch Wahrscheinlichkeiten ausgedrückt. Es steht I für das Ereignis „infiziert“, und T für das Ereignis „Testergebnis positiv“.

Prävalenz ist die Wahrscheinlichkeit $P(I)$ für I in der getesteten Population, hier ausgedrückt durch die relative Häufigkeit $(a + c)/n$. Falls die Testpersonen einen zufälligen Teil der Gesamtpopulation repräsentieren, ist dies ungefähr gleich der Prävalenz p in der Gesamtbevölkerung, also $p \approx (a + c)/n$.[9]

Sensitivität ist die bedingte Wahrscheinlichkeit $\sigma = P(T|I)$ eines positiven Testergebnisses unter der Voraussetzung, dass die Infektion vorliegt, hier ausgedrückt durch die relative Häufigkeit $a/(a + c) \approx \sigma$.[10]

Spezifität ist die bedingte Wahrscheinlichkeit[11] $\tau = P(\neg T|\neg I)$ eines negativen Testergebnisses unter der Voraussetzung, dass keine Infektion vorliegt, hier ausgedrückt durch die relative Häufigkeit $d/(b + d) \approx \tau$.[12]

Vorhersagewert ist die bedingte Wahrscheinlichkeit $\psi = P(I|T)$ dafür, dass bei positivem Testergebnis tatsächlich die Infektion vorliegt. Er wird hier ausgedrückt durch die relative Häufigkeit $a/(a + b)$. Der komplementäre Wert $\varphi = 1 - \psi = P(\neg I|T)$ ist die **Fehlalarmrate**, also $\varphi \approx b/(a + b)$.

Der Vorhersagewert ist also die Wahrscheinlichkeit für das Zutreffen des Umkehrschlusses von der Wirkung auf die Ursache. Die Formel von Bayes drückt ihn durch die übrigen Größen aus:

$$P(I|T) = \frac{P(T|I) \cdot P(I)}{P(T)} \quad \text{bzw.} \quad \psi = \frac{\sigma \cdot p}{P(T)}\,. \tag{1}$$

Das ist kein tiefliegender mathematischer Satz, sondern folgt direkt aus der Definition der bedingten Wahrscheinlichkeit:

$$P(T|I) = \frac{P(T \cap I)}{P(I)} \quad \text{und} \quad P(I|T) = \frac{P(T \cap I)}{P(T)}\,,$$

[9] Das „Näherungsweise-gleich“-Zeichen $\approx$ steht hier und im Folgenden, weil die theoretischen Werte $p, \sigma, \ldots$ durch die in der Stichprobe beobachteten Werte $(a + c)/n$, $a/(a + c)$, $\ldots$ angenähert werden, aber nicht unbedingt genau mit ihnen übereinstimmen.

[10] Der komplementäre Wert $\beta = 1 - \sigma = P(\neg T|I)$ wird auch Fehler 2. Art genannt: Eine vorhandene Infektion wird nicht erkannt.

[11] Das Symbol $\neg$ bedeutet die logische Negation.

[12] Der komplementäre Wert $\alpha = 1 - \tau = P(T|\neg I)$ wird auch Fehler 1. Art genannt: Eine nicht vorhandene Infektion wird fälschlicherweise behauptet.

wobei $P(T \cap I)$ die Wahrscheinlichkeit dafür ist, dass T und I gemeinsam auftreten.

Schöner wäre es, wenn auf der rechten Seite der Formel (1) nur die Größen Prävalenz p, Sensitivität σ und Spezifität τ aufträten. Das kann man leicht erreichen, indem man die „störende" Wahrscheinlichkeit $P(T)$ für ein positives Testergebnis durch diese Größen ausdrückt: Da T die disjunkte Vereinigung $(T \cap I) \cup (T \cap \neg I)$ ist, ist

$$P(T) = P(T|I) \cdot P(I) + P(T|\neg I) \cdot P(\neg I) = \sigma \cdot p + (1-\tau)(1-p),$$

wobei $P(\neg I) = 1 - P(I)$ und $P(T|\neg I) = 1 - P(\neg T|\neg I)$. Setzt man dies in die Formel (1) ein, so erhält man eine alternative Formel, auf deren rechter Seite nur noch die erwünschten Größen auftreten[13]:

$$P(I|T) = \frac{P(T|I) \cdot P(I)}{P(T|I) \cdot P(I) + (1 - P(\neg T|\neg I)) \cdot (1 - P(I))}\,, \tag{2}$$

oder mit den Kurzbezeichnungen ausgedrückt:

$$\psi = \frac{\sigma \cdot p}{\sigma \cdot p + (1-\tau) \cdot (1-p)}\,, \tag{3}$$

an der man die Abhängigkeit von ψ von den anderen drei Größen p, σ und τ ablesen kann. Die komplementäre Formel für die Fehlalarmrate $\varphi = 1 - \psi$ ist

$$\varphi = \frac{(1-\tau) \cdot (1-p)}{\sigma \cdot p + (1-\tau) \cdot (1-p)}\,. \tag{4}$$

Als Beispiel betrachten wir ψ als Funktion von p bei festen σ und τ, also die funktionale Abhängigkeit des Vorhersagewerts von der Prävalenz für einen Test mit bekannter Sensitivität und Spezifität.
Die Abbildung 3 zeigt diese Funktion für das Beispiel mit $\sigma = 98\,\%$ und $\tau = 95\,\%$. Man erkennt, dass ψ als Funktion von p zwischen 0 und 1 monoton von 0 bis 1 zunimmt. Ab einer Prävalenz von ungefähr 0.2, also $20\,\%$, liegt der Vorhersagewert mindestens bei 0.8, also $80\,\%$, was man als einigermaßen akzeptabel ansehen würde. In der Praxis ist die Prävalenz glücklicherweise meistens sehr viel niedriger, was allerdings *den unerwünschten, aber unvermeidbaren Nebeneffekt eines niedrigen bis miserabel niedrigen Vorhersagewerts mit sich bringt.*

---

[13] Man kann die Richtigkeit der Formel auch anhand der relativen Häufigkeiten nachvollziehen: In der Formel steht dann $\frac{a}{a+c}$ für die Sensitivität $P(T|I)$ und $\frac{a+c}{n}$ für die Prävalenz $P(I)$, also $\frac{a}{n}$ für das Produkt $P(T|I) \cdot P(I)$, das sowohl im Zähler als auch im Nenner steht. Im Nenner kommt dann noch die Spezifität $P(\neg T|\neg I)$ vor, die durch $\frac{d}{b+d}$ repräsentiert wird. Damit wird $1 - P(\neg T|\neg I)$ zu $1 - \frac{d}{b+d} = \frac{b}{b+d}$, und weil $1 - P(I)$ zu $1 - \frac{a+c}{n} = \frac{b+d}{n}$ wird, wird das Produkt $(1 - P(\neg T|\neg I)) \cdot (1 - P(I))$ im Nenner zu $\frac{b}{b+d} \cdot \frac{b+d}{n} = \frac{b}{n}$. Der gesamte Nenner wird also zu $\frac{a}{n} + \frac{b}{n} = \frac{a+b}{n}$, der ganze Bruch also zu $\frac{a}{n} / \frac{a+b}{n} = \frac{a}{a+b}$, was ja gerade für den Vorhersagewert steht.

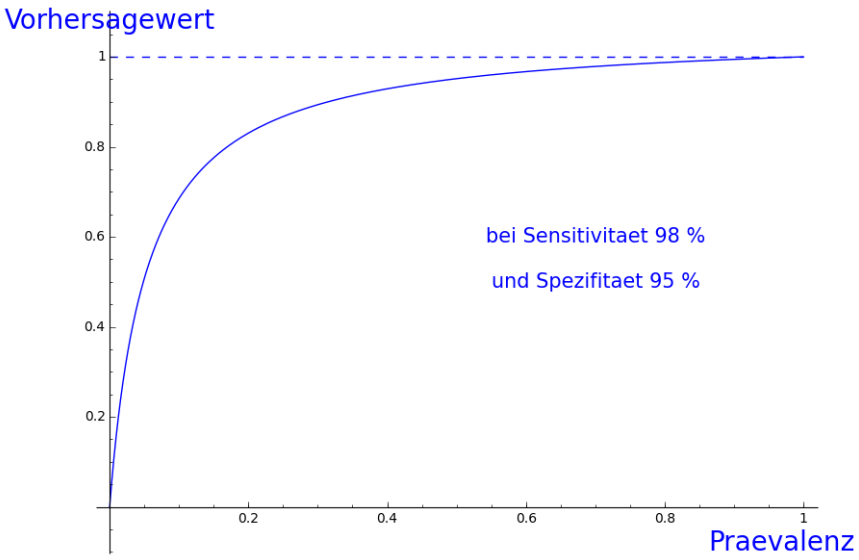

Abbildung 3: Der Vorhersagewert als Funktion der Prävalenz bei fester Sensitivität und Spezifität

Andere Testszenarien

Das beschriebene Testszenario läßt sich auf andere Situationen übertragen – immer dann, wenn es darum geht, aus einer großen Menge die Individuen mit bestimmten Eigenschaften herauszufischen. Je nach konkretem Anwendungsfall möchte man unterschiedliche Parameter optimieren.

Materialprüfung

Werden Geräte oder Materialien, die in den Handel gebracht werden sollen, getestet, darf *kein einziges* davon unsicher sein. Das bedeutet: Der Test muss eine Sensitivität von 100 % haben – jedes defekte Stück muss zuverlässig erkannt und aussortiert werden; dass dabei auch einige „gute“ durchfallen, nimmt man mehr oder weniger zähneknirschend in Kauf. D. h. bei der Abwägung zwischen Sensitivität und Spezifität hat die Sensitivität höchste Priorität.

Die theoretische Analyse führt im Prinzip auf die gleichen Probleme wie bei medizinischen Tests: Die unvermeidliche Fehlalarmrate ist hoch, hier aber nicht so folgenschwer. Aber sie ist in diesem Kontext vermutlich auch deutlich geringer, denn auf der praktischen Seite stellt sich die Situation um einiges günstiger dar: Im Gegensatz zu medizinischen Tests sind in einer technischen Umgebung

- die Messungen in der Regel viel genauer,
- die Verfahrensabläufe besser kontrollierbar,
- der Stress, und damit die Gefahr menschlicher Fehler, geringer.

Klinische Prüfungen

Bei einer klinischen Prüfung geht es darum, die Wirksamkeit eines Medikaments (oder einer anderen Therapieform) zu beurteilen. Hier werden vorrangig die Größen

- Fehler 1. Art $\alpha = 1 - \tau$, also die Behauptung eines Therapieerfolgs einer eigentlich unwirksamen Therapie,
- Fehler 2. Art $\beta = 1 - \sigma$, also die Ablehnung einer eigentlich wirksamen Therapie,

ins Visier genommen. Für α setzt man ein Limit, meistens 5 % oder 1 %; dann wird unter allen statistischen Tests der mit der höchsten „Power" = Sensitivität σ gesucht. Der Vorhersagewert bzw. die Fehlalarmrate sind in diesem Szenario keine sinnvollen Größen: Auf welche Grundgesamtheit sollten sie sich denn beziehen?

Informationssuche

Bei der einer Informationssuche, etwa mit einer Internet-Suchmaschine oder einem Literatur-Recherchesystem, stellt die Nutzerin eine Anfrage der Art: „Finde alle Webseiten, die Informationen über CoViD-19 enthalten." Als Ergebnis wird sie eine (in diesem Fall sehr lange) Liste von Treffern erhalten.[14] Viele davon enthalten relevante Information, viele nicht (z. B. wenn sie den Suchbegriff nur nebenbei erwähnen oder „alternative Fakten" präsentieren). Zum Verständnis hilft wieder die Vierfeldertafel, passend beschriftet:

gefunden („Treffer")	relevant? ja	nein	Summe
ja	a	b	$a+b$
nein	c	d	$c+d$
Summe	$a+c$	$b+d$	n

Von den insgesamt durchsuchten n Webseiten sind $a+c$ relevant und $b+d$ irrelevant.[15] Von den relevanten sind a in der Trefferliste enthalten, aber c nicht. Dafür enthält die Trefferliste auch b vermeintliche Treffer, die nicht relevant sind.

In diesem Kontext ist es sinnvoll, die Fehlalarmrate zu minimieren, denn sie gibt an, wie viele von den Treffern irrelevant sind – wer will sich schon durch riesige Mengen irrelevanter Texte quälen? Es reicht ja, ein paar wirklich relevante zu finden, denn in den meisten relevanten wird wohl mehr oder weniger die gleiche Information stecken. Welche Freiheitsgrade für diese Minimierung gibt es? Nun, die Gesamtzahl n liegt fest (zumindest im Moment der Suche), ebenso die Prävalenz p (siehe Fußnote 15). An Formel 4 sieht man direkt, dass die Fehlalarmrate φ bei wachsender Sensitivität σ sinkt, aber die Sensitivität soll hier ja nicht so wichtig sein. Und Formel 3 zeigt, dass der Vorhersagewert ψ mit wachsender Spezifität τ steigt, die Fehlalarmquote also ebenfalls sinkt. Das wenig überraschende Fazit aus dieser Betrachtung ist, dass der Algorithmus, mit dem die Relevanz bewertet wird, die Spezifität maximieren sollte.

[14] die sinnvollerweise nicht alle direkt angezeigt werden, siehe unten

[15] unter der Annahme, dass es überhaupt ein irgendwie sinnvolles Kriterium dafür gibt

Ein praktischer Ansatz ist der Einsatz eines Algorithmus, der die Relevanz quantifiziert, also einen „Relevanz-Score“ berechnet. Damit wird die Trefferliste nach Relevanz sortiert und dann nach einer gewissen Zahl von besonders hoch bewerteten Treffern abgeschnitten. Das bedeutet, dass die Anforderungen an die Relevanz sehr hoch geschraubt werden. Im Idealfall ist dann $b = 0$, also die Spezifität $\tau = 1$ und die Fehlalarmrate $\varphi = 0$ – die Sensitivität σ spielt dann für die Optimierung der Fehlalarmquote keine Rolle mehr. Sie leidet aber potenziell stark unter dem „Abschneide-Algorithmus“, und das hat die unvermeidliche Nebenwirkung, dass sehr viele relevante Webseiten nicht gefunden werden. Zwar wollten wir das ja in Kauf nehmen – unter der Nebenbedingung „Fehlalarmrate 0“ kann man dann aber doch versuchen, die Sensitivität zu optimieren. Wichtig ist in diesem Kontext natürlich, dass der Bewertungs-Algorithmus Falschinformationen erkennt und abwertet und die wirklich relevanten Treffer nicht versehentlich zu gering bewertet. Aber das ist ein anderes Thema.

Anhang: Wieviel Infizierte gibt es tatsächlich?

Der aktuelle Stand der SARS-CoV-2-Pandemie Mitte November 2020

Kurze Antwort für die aktuelle Pandemie: *Das weiß niemand*. Bekannt sind nur die Zahlen der positiven Testergebnisse und die Zahl der tatsächlich mit handfesten Symptomen Erkrankten mit gesicherter Diagnose oder gar in Intensiv-Behandlung. Wenn die Diagnose gesichert ist, wird der jeweilige Fall natürlich bei den positiv Getesteten mitgezählt.

Viel Verwirrung entsteht dadurch, dass in der öffentlichen Diskussion mehrere unterschiedliche Größen munter vermischt und z. T. fehlinterpretiert werden. Abbildung 4 illustriert:

1. die Zahl $a + b$ der positiv auf SARS-CoV-2 Getesteten (a = die wirklich Infizierten darunter, b = die falsch Positiven),
2. die Zahl der tatsächlich mit SARS-CoV-2 Infizierten, die niemand kennt,
3. die Zahl n_k der an SARS-CoV-2 Erkrankten (also CoViD-19-Fälle),
4. die Zahl n_i der CoViD-19-Fälle auf Intensivstationen.

Die Zahl, die vor allem Anlass zu ernster Sorge gibt, ist die Zahl n_i unter Punkt 4. Die Zahl, die in der öffentlichen Wahrnehmung an erster Stelle genannt wird, ist die Zahl $a + b$ unter Punkt 1. Diese steigt in der Tat[16] in den letzten Wochen (fast) weltweit enorm. Allerdings kann diese Anzahl der positiv Getesteten in die Irre führen, das ist das in diesem Artikel analysierte Fehlalarm-Paradoxon: *Selbst noch so gute Tests produzieren* **unvermeidlich** *viele Fehlalarme*, d. h., die Zahl b der Fehlalarme ist ziemlich groß, aber eben auch unbekannt.

[16] laut Robert-Koch-Institut und anderen hinreichend zuverlässigen Quellen

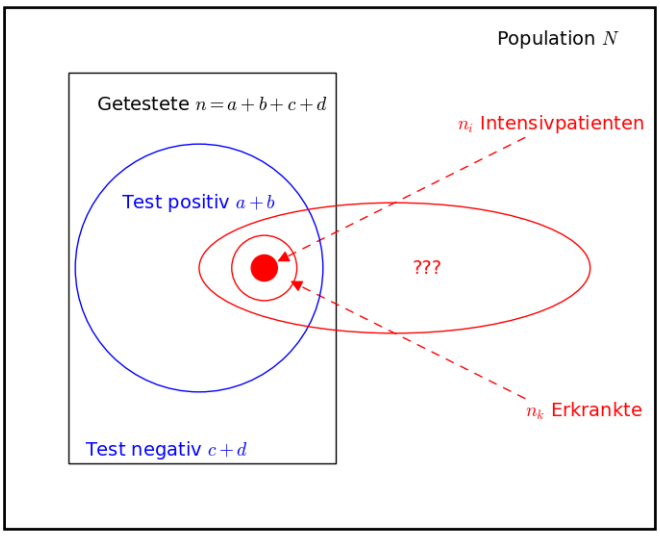

Abbildung 4: Die zu unterscheidenden Anzahlen von positiv Getesteten, Erkrankten und Intensivpatienten; die drei „???“ stehen für die unbekannte Anzahl der Infizierten.

Die Anzahl a der tatsächlich Infizierten unter den positiv Getesteten liegt mit Sicherheit sehr viel niedriger (20 % wäre, wie wir gesehen haben, möglicherweise schon hoch gegriffen, aber de facto ist diese Zahl auch unbekannt und kann nur sehr grob geschätzt werden). Die Zahl der falsch Negativen kann man wohl als vernachlässigbar ansehen. Andererseits kommen aber die unbekannten Infizierten dazu, die nicht getestet wurden, z. B. weil sie symptomfrei[17] sind. D. h., die Zahl der aktuell tatsächlich Infizierten unter Punkt 2 liegt als Summe zweier Dunkelziffern selbst im Dunkeln. Man darf allerdings vermuten, dass sie z. Z. *ebenfalls rasant ansteigt*.

Begründung: Bei gleichem Testverfahren bleiben Sensitivität σ und Spezifität τ konstant. Selbst wenn man die Prävalenz p, und damit nach Gleichung (3) auch den Vorhersagewert ψ, als konstant annimmt, steigt die Zahl $b \approx \psi \cdot (a + b)$ proportional zur Zahl $a + b$ der positiv Getesteten. Wächst[18] p, so nach Gleichung (3) bzw. Abbildung 3 auch ψ. Und damit wächst b *sogar überproportional* mit $a + b$. Nicht berücksichtigt ist hier die Dunkelziffer der Infizierten, aber nicht Getesteten, die natürlich einige Unsicherheit in diese Abschätzung bringt.

Andererseits kann man vermuten, dass die Zahl n_k unter Punkt 3 ziemlich exakt ist, denn hierbei handelt es sich ja um gesicherte Diagnosen.

Vieles, insbesondere die Zahl der tatsächlich Infizierten, ist leider (unvermeidbar aufgrund der Datenlage) nur ziemlich ungenau schätzbar. *Am besten schaut man daher auf die Anzahl n_i der Fälle, die in den Intensivstationen behandelt werden* (Punkt 4). Das ist ein hartes Faktum und eigentliche Begründung für die derzeitigen Einschränkungen, die eine Überlastung des

[17] aber nach derzeitiger Erkenntnis oft trotzdem ansteckend

[18] und das dürfen wir getrost annehmen, wenn die Zahl der positiven Tests stark steigt, jedenfalls stärker, als nur allein durch die Ausweitung der Testkapazitäten zu erklären

Gesundheitssystems verhindern sollen, wie sie sich in Nachbarstaaten – z. B. in einigen Schweizer Kantonen – schon wieder abzeichnet.

Zusammenfassung der Kernaussagen

Obwohl die ungenaue Datenlage und ungesicherte Datenqualität keine exakten Aussagen oder gar Vorhersagen zulassen, kann man durchaus einige qualitative Schlüsse ziehen:

- Die Zahl der positiv Getesteten stellt eine deutliche Überschätzung der Zahl der tatsächlich Infizierten dar, die meisten der positiven Tests sind Fehlalarme.
- Ein positives Testergebnis ist noch weit von einer endgültigen Diagnose entfernt. Es muss durch weitere Untersuchungen überprüft werden, um zu einer gesicherten Diagnose zu kommen.
- Wird nur ein vorgesiebter Personenkreis getestet (z. B. Leute, bei denen einschlägige Krankheitssymptome beobachtet wurden), so kann man von einer mehr oder weniger erhöhten Prävalenz in dieser Gruppe und somit von einer deutlich geringeren Fehlalarmrate ausgehen.
- Negativ Getestete sind mit hoher Sicherheit nicht infiziert.
- Positiv Getestete, bei denen ein zweiter Test negativ ausfällt, sind mit ziemlich hoher Sicherheit nicht infiziert.
- Die Zahl der tatsächlich Infizierten ist unbekannt.
- Dennoch kann man sagen, dass die Zahl der tatsächlich Infizierten überproportional mit der Zahl der positiv Getesteten zunimmt.
- Die starke Zunahme der positiven Testergebnisse in den Monaten September bis November 2020 (bei in etwa gleichbleibender Zahl von täglichen Tests) deutet auf eine sehr starke Zunahme der Zahl an tatsächlich Infizierten hin.
- Für die Entscheidung über notwendige Maßnahmen zur Eindämmung der Pandemie-Folgen schaut man am besten auf die gesicherte Anzahl der Fälle, die in den Intensivstationen behandelt werden, denn eine Überlastung hier ist ein nicht tragbares Risiko für das Gesundheitssystem.