

Diagnostische Forschung – den Blickwinkel erweitern

Bericht von einem Workshop veranstaltet von der Abteilung für Allgemeinmedizin, Philipps-Universität Marburg

Johannes Hauswaldt

Hintergrund

Für ihre diagnostischen und therapeutischen Entscheidungen verfügen Ärztinnen und Ärzte über vielfältige Informationen: den unmittelbaren Eindruck vom Patienten, Vorgeschichte und Befund, Schnelltests und Laboruntersuchungen, Bildgebung.

Klassische diagnostische Studien im Vierfelder-Design reichen oft nicht aus, die klinische Vielfältigkeit in diesem Prozess angemessen abzubilden. Vereinfachende Herangehensweisen (*fast and frugal heuristics*) auf komplexem Hintergrund sind vermutlich besser geeignet, den Entscheidungsprozess von Arzt und Patient (dichotomisieren, individualisieren, Werte berücksichtigen) widerzuspiegeln.

Hier können klinische Epidemiologie, Versorgungsforschung, statistische Methoden, Kognitions- und Informationswissenschaften gemeinsam den Blick auf die medizinische Entscheidungsfindung erweitern – vorausgesetzt, dass die Forscher aus diesen unterschiedlichen Wissenschaftsfeldern zu einer gemeinsamen Sprache finden.

Bericht

Zum Workshop „*Diagnostic Research – Broadening the Perspective*“ am 13. und 14. März 2013 begrüßte Erika Baum, Marburg, 28 Teilnehmer aus sieben Ländern.

Norbert Donner-Banzhoff, Marburg, führte in den Workshop ein (Abb. 1) und referierte anstelle des erkrankten Ralph Hertwig zum Thema „Einfache Heuristiken für komplexe Probleme“ („*Simple heuristics for complex problems*“). Es gebe

zwei Ansichten darüber, was eine rationale Vorgehensweise sei:

- eine, die sich logischer Regeln, statistischer Untersuchungen und Erwartungen über den möglichen Nutzen bedient, alle möglichen Informationen sammelt, ausgeklügelte Modelle anwendet, optimiert und möglicherweise auch übermäßig anpasst,
- und eine, die auch das Umfeld der Entscheidung mit berücksichtigt, die meisten Informationen ignoriert und der oftmals ein einziger Grund für eine Entscheidungsfindung genügt, die mit „Weniger ist mehr“ zufrieden und ökologisch adaptiert ist.

Exponenten der beiden Denkweisen sind einerseits der Nobelpreisträger Daniel Kahnemann, andererseits Gerd Gigerenzer und Gary Klein. Der Marburger Herz-Score ist ein Beispiel für eine derartige abwägende Heuristik.

In der lebhaften Diskussion wurden eine Reihe von Gegensätzen genannt: klinische gegenüber forschender Herangehensweise; beschreibende versus normensetzende Forschung; „Information“ und „Gestalt“ (medizinische Probleme sind meist einfach, können gelegentlich aber auch sehr schwierig und komplex sein). Geert-Jan Dinant, Maastricht, wies darauf hin, dass gerade Generalisten nicht selten von den Befunden direkt zu Behandlung und Prognose übergehen. Dagegen wurde eingewandt: „Intuition liegt meistens richtig; liegt sie aber falsch, dann auch fürchterlich falsch“ und „Die meisten Modelle sind zwar falsch, aber einige können doch nützlich sein“ (soll von Niels Bohr stammen; wird aber auch Lord Rutherford oder George Box zugesprochen).

Karel Moons, Utrecht, mit dem Thema „Wissenschaftliche Bewertung diagnostischer und prognostischer Biomarker, Tests und Modelle“ („*Scientific evaluation of diagnostic and prognostic (bio)markers, tests and models – from practice to research and back*“) stellte die schnell wachsende Zahl diagnostischer Tests und Biomarker infrage, weil er starke Zweifel hegt, dass sie alle klinisch relevant seien. Keine Diagnose oder Prognose werde durch einen einzelnen Test oder Marker allein gestellt. Für Anwender sei der eventuelle Zusatznutzen gegenüber dem bis dahin erlangten Wissen wichtig, auch dieser müsse daher erforscht und quantifiziert werden: Viele Tests mögen für sich genommen aussagekräftig sein – in ihrer Kombination sind sie es nicht.

Moons schlug vor, neue Tests und Biomarker in Stufen zu evaluieren:

- zunächst einmal für sich,
- dann in Bezug auf den Zusatznutzen gegenüber vorhandenen und bewährten Verfahren,
- auch in Bezug auf die Konstruktion des zugrunde liegenden Vorhersagemodells,
- zuletzt über eine Bewertung der Wirksamkeit des Vorhersagemodells.

Kein Test sollte ohne externe Validierung angewandt werden und am besten ausschließlich bei Nachweis von Wirksamkeit/Zusatznutzen. Ärzte wüssten mit Vorhersagemodellen nicht viel anzufangen, weil sie in deren Anwendung keine Routine hätten.

In der Diskussion wurde die Problematik von „Sensitivität“ und „Spezifität“ als keineswegs konstante Testeigenschaften herausgestellt. Diese Testgütekriterien seien (außer in einer frühen

Phase der Markerbewertung) kaum hilfreich, sagen nichts über den Zusatznutzen aus, auch führe die erforderliche Dichotomisierung der Markerergebnisse zu Informationsverlust. Starke ökonomische Interessen würden hier das Feld von Forschung und Veröffentlichungen bestimmen, mit Reaktionen in verschiedenen „Gesundheitsmärkten“ (beispielsweise Verweigerung von Kostenrückerstattung oder staatliche Regulierung). Patienten hätten ein Interesse an weniger invasiven Maßnahmen und Kosteneffektivität.

Eike Hüllermeier, Marburg, schloss mit seinem Vortrag „Maschinenlernen für die medizinische Diagnosestellung“ („Machine learning for medical diagnosis“)

an. Während klassische Expertensysteme enzyklopädisches Wissen elektronisch sammeln, seien künstliche neuronale Netzwerke zwar faszinierend, letztlich jedoch in ihrer inneren Struktur unverständlich und nicht repräsentativ. Modernes maschinelles Lernen sollte Modelle erzeugen, die verständlich sind und neue Einsichten erlauben. Es gehe letztlich darum, Maschinen zu erschaffen, die lernen können: autonome Computersysteme, die in der Lage sind, ihre Leistung auf der Basis von Erfahrung zu verbessern, um die klinische Entscheidungsfindung zu unterstützen, und die sich an veränderte Umgebungsbedingungen anpassen.

Hüllermeier beschrieb verschiedene mathematische Verfahren des Klassifizierens, als Grundlage für dichotome Entscheidungen und erweiterte diese für polychotome Ausgangslagen. Als willkommenes „Nebenprodukt“ der mathematischen Formulierung werden zwei Arten von Unsicherheit unterscheidbar, nämlich epistemische Unsicherheit in der Folge von Nicht-Wissen, also Mangel an Information und durch mehr Information behebbar, gegenüber aleatorischer Unsicherheit als Folge von „Rauschen in den Daten“. Für die medizinische Diagnosestellung werde Automatenlernen zunehmend wichtig: Damit können Gewichtung von Einflussfaktoren und deren Interaktionen auf nicht-additive Weise zusammengefasst und operationalisiert werden.

In der Diskussion wurde auf den Bias durch (fehlenden) Einschluss (*inclusion/incorporation bias*) eingegangen: Hintergrundwissen sei vermutlich die Grundlage für gelingendes Maschinenlernen. Unsicherheit kontrastiere hier mit dem Unerwarteten.

Am zweiten Tag des Workshops führte Benjamin Djulbegovic, Tampa/USA, den Aspekt des Bedauerns, der Reue (*regret*) in das Feld medizinischer Entscheidungsfindung ein („Diagnostic testing in medicine: regret approach“). Weil Entscheidungen kategorial (dichotom, polychotom) sein müssen, sei das Konzept von „Schwellen“ (*threshold concept*) ein Dreh- und Angelpunkt bei der Entscheidung über die Ausdehnung des diagnostischen Prozesses (Abb. 2). Gedankliche Vorwegnahme von Handlungsfolgen, also die Fähigkeit des menschlichen Vorstellungsvermögens zum kontrafaktischen Denken lägen dem „Reue-Ansatz“ zugrunde. Weil irrtumsfreie medizinische Empfehlungen nicht möglich seien, gebe es einen Bereich akzeptierten Bedauerns und Toleranz gegenüber diagnostischen Fehlern. Djulbegovic schloss seinen Vortrag mit der Analyse von werte- und bedauernsbasierten Entscheidungskurven sowie von visuellen Analogskalen zur Messung von *regret* ab.

Die lebhafteste Diskussion behandelte zunächst die Gründe für das Übermaß an überflüssigen diagnostischen Tests: Sind es kommerzielle Interessen oder angenommene Bedrohung durch Kunstfehlerprozesse? Möglicherweise, so Djulbegovic, könne anstelle isolierter

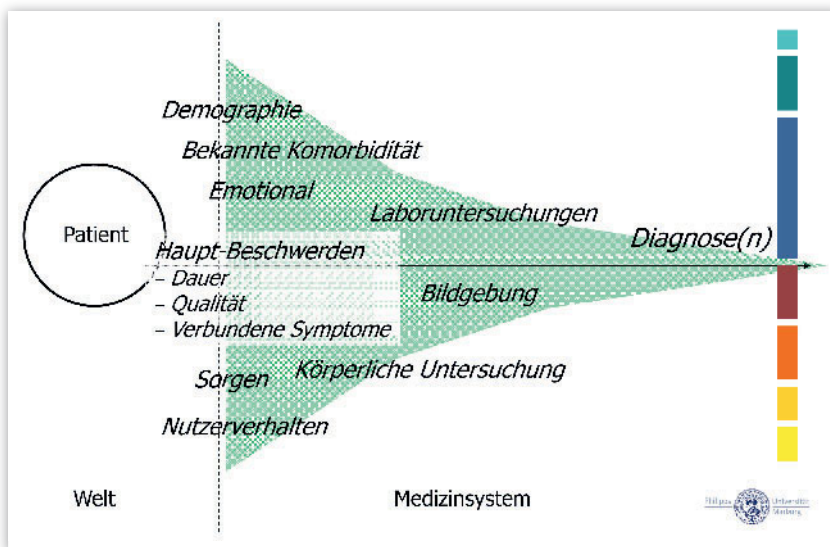


Abbildung 1 Diagnostizieren – Unsicherheit verringern (N. Donner-Banzhoff, modif.)

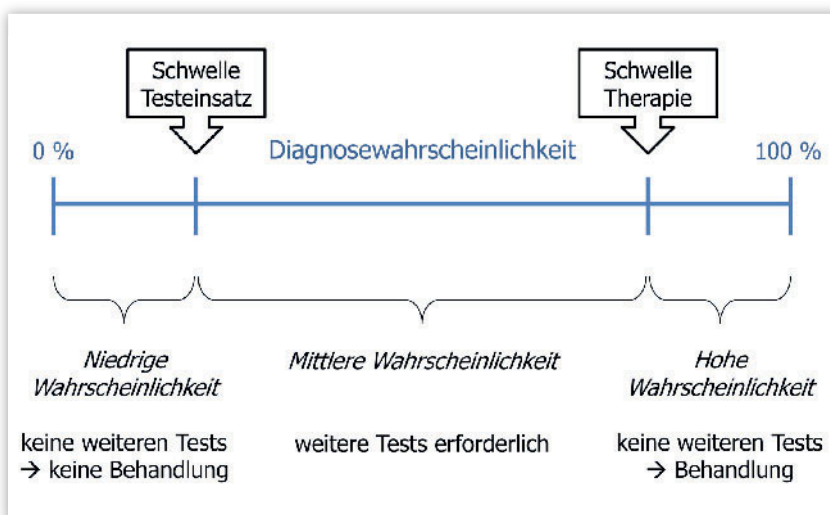


Abbildung 2 Kontinuum der Diagnosewahrscheinlichkeit und das Schwellenkonzept (B. Djulbegovic, modif.)

Tests ein umfassender Ansatz hilfreich sein: „Ich bin zu 80 % überzeugt, dass ich Ihnen helfen kann – sind Sie bereit, die restlichen 20 % zu akzeptieren?“ Mit visuellen Analogskalen, so Donner-Banzhoff, könnten mehrere Bereiche, die anders nicht direkt miteinander vergleichbar sind, zusammengefasst werden. Auch könne man mit ihnen Vorstellungen von nachträglicher Reue „vorwegnehmen“; jedoch müssten Patienten und Ärzte ihre Anwendung noch lernen. Bezüglich des Anspruchs 100 %iger Sicherheit in medizinischer Diagnostik oder Behandlung sei ein kultureller und politischer Wandel erforderlich. Partizipative Entscheidungsfindung (*shared decision making*) müsse auch auf diagnostische Prozesse angewandt werden. Ärzte sollten aus der Psychologie lernen, wie Wahrscheinlichkeiten kommuniziert werden können.

Jörg Haasenritter und Norbert Donner-Banzhoff, Marburg, referierten anschließend über ein umfassendes Design für diagnostische Studien („*The comprehensive diagnostic study (CDS) design*“) – einen von ihnen entwickelten systematischen Ansatz, um Untersuchungen, die eine einzelne spezifische Diagnose sicherer bestätigen oder ausschließen, mit Studien zu verbinden, die Symptome während der differenzierenden Ent-

scheidung über mehrere Diagnosen bewerten. Schlüsselgedanke dabei sei die Quantifizierung der Informationsdichte (der Entropie nach Shannons Informationstheorie), die als Maß für Unsicherheit im diagnostischen Prozess genommen werde. Verschiedene mögliche Herangehensweisen in diesem Design demonstrierten die Vortragenden anhand eigener Untersuchungsergebnisse bei Patienten mit Brustschmerz.

Die Diskussion fragte zunächst danach, was geschieht, wenn sich die „Natur von Krankheiten“ nicht an die ausgewählten Testkriterien halte? Was, wenn Kategorien einander überschneiden? Wie verhält es sich bei unterschiedlichen Intentionen während der medizinischen Entscheidungsfindung, etwa Gefahrenabwehr (AGV) oder Filterfunktion gegenüber dem Festlegen einer Diagnose? Was ist der Einfluss von Rahmenbedingungen, etwa der Reihenfolge der gestellten diagnostischen Fragen? Dichotomisierende Entscheidungsbäume sind zum Ausschluss schwerwiegender Probleme oder für kritische Entscheidungen vermutlich hilfreich; Versuche, die stattfindenden Prozesse zu quantifizieren und mathematisch zu modellieren, könnten nützlich sein, um die schnell wachsende Komplexität bei Einschluss vieler Diagnosen zu vereinfachen.

Im abschließenden, großzügig bemessenen Zeitabschnitt des Workshops trugen alle Teilnehmer in Form eines kollektiven Brainstorming weitere Resultate zusammen und entwickelten Wünsche und Planungen für gemeinsame Schritte, um ärztliches Diagnostizieren breiter zu erforschen. Sie begrüßten den Vorschlag, die Ergebnisse des Workshops in einer Deklaration zusammenzufassen, wozu sich Norbert Donner-Banzhoff, Marburg, Geert-Jan Dinant, Maastricht/Niederlande, und Jean-Marie Degryse, Löwen/Belgien, bereit erklärten.

Danksagung an Norbert Donner-Banzhoff für die kritische Durchsicht des Manuskripts und Anmerkungen

Korrespondenzadresse

Dr. med. Johannes Hauswaldt, MPH
 Institut für Allgemeinmedizin,
 Universitätsmedizin Göttingen (UMG)
 Humboldtallee 38, 37073 Göttingen
 Tel.: 0551 3919942
 Johannes.Hauswaldt@med.uni-goettingen.de
<http://www.allgemeinmedizin.med.uni-goettingen.de>