Evidence-based Medicine: Wie beurteile ich eine Studie zu einer therapeutischen Intervention?*

D. Pewsner[a], P. Jüni[b], H. C. Bucher[c]

Ärztliche Fortbildung zielt wesentlich auf die Verbesserung bestehender Therapieansätze. Nicht zuletzt geschieht dies durch die pharmazeutische Industrie, welche durch Artikel, Informationsbroschüren und neuerdings auch im Internet Grundversorger und Kliniker förmlich überflutet. Durch die Beachtung einiger weniger Kriterien in der Beurteilung der Validität von Studien kann erlernt werden, relevante Information von fragwürdigen Forschungsergebnissen zu unterscheiden. Es ist ein Anliegen von «Evidence-based Medicine» (EBM), Grundversorger und Kliniker zu befähigen, Aussagen von Studien kritisch zu beurteilen und zu interpretieren. Anhand eines Beispiels möchten wir aufzeigen, wie eine Therapiestudie beurteilt werden kann. Wir stützen uns dabei weitgehend auf die Vorschläge von Guyatt et al. [1, 2].

Klinisches Szenario

Sie betreuen als Hausarzt eine 48jährige Patientin, die infolge von Eheproblemen unter einer mässig schweren reaktiven Depression leidet, welche schon mehrere Wochen andauert. Neben den für diese Erkrankung typischen Symptomen besteht eine ausgeprägte Ängstlichkeit. Auf Ihren Vorschlag einer antidepressiven Pharmakotherapie reagiert die Patientin abwehrend. Allerdings berichtet sie über einen vor kurzem gelesenen Artikel über Johanniskraut bei Depressionen. Allenfalls wäre sie bereit, sich mit einem «nicht-chemischen» Phytotherapeutikum behandeln zu lassen. Auch Sie haben schon Artikel über Hyperikum bei Depressionen gelesen, die Sie jedoch nicht überzeugt haben. Sie versprechen Ihrer Patientin, mittels einer vertieften Literatursuche mehr über diese therapeutische Alternative herauszufinden.

Die Evidenz suchen

Ihr Plan ist es, Information über den Stellenwert von Hypericum bei der Therapie von Depressionen zu finden. Sie sind mit «Pubmed» vertraut und suchen mit dem Suchsystem «MeSH-Browser» (Medical Subject Headings) den indexierten Suchbegriff «Hypericum». Als zweiten MeSH-Begriff geben Sie «Depression» ein. Da Sie primär an Studien zu Therapiemassnahmen mit der höchsten Evidenz interessiert sind, suchen Sie randomisierte kontrollierte Studien und schränken Ihre Suche mit dem Publikationstyp «randomised controlled trial» ein. Die Kombination der beiden MeSH-Suchbegriffe mit dem Publikationstyp ergibt für den Zeitraum 1966 bis 2000 insgesamt 5 Arbeiten. Bei der Durchsicht finden Sie zwei Arbeiten, die Sie interessieren. Bei beiden handelt es sich um randomisierte kontrollierte Studien; in der einen wird Hyperikum mit einem selektiven Serotonin-Wiederaufnahmehemmer (SSRI) Sertaline (Gladem®, Zoloft®), in der andern Hyperikum (Hyperikum Extrakt ZE 117, Remotiv®) 2×250 mg mit dem Trizyklikum Imipramin 2×75 mg tgl. (Tofranil®) verglichen. Da es sich bei der ersten Arbeit um eine Pilotstudie mit nur 30 Patienten handelt, entscheiden Sie sich für die über 10mal grössere Studie aus dem British Medical Journal [3]. Das British Medical Journal ist im «Fulltext» frei erhältlich, was Ihnen von

* Für die Diagnostik, siehe Swiss Medical Forum Nr. 9 vom 28.2.01, S. 213-20.

[a] Praxis für Innere Medizin FMH, Bern
[b] Rheumatologische Universitätsklinik, Inselspital Bern
[c] Medizinische Universitäts-Poliklinik, Kantonsspital Basel

Korrespondenz:
PD Dr. Heiner C. Bucher
Medizinische Universitäts-Poliklinik
Kantonsspital
CH-4031 Basel

hbucher@uhbs.ch

«Pubmed» (http://www.ncbi.nlm.nih.gov) aus den direkten Zugang zu dieser Arbeit ermöglicht.

Kritische Beurteilung und Interpretation der Studie

Bei der Beurteilung einer Studie zu einer therapeutischen Intervention stellen wir uns drei Hauptfragen:

1. Sind die Ergebnisse der Therapiestudie valide, d.h., sind die methodische Ausführung korrekt und die Ergebnisse frei von Verzerrungen?
2. Sind die Ergebnisse der Studie klinisch bedeutsam?
3. Sind die Ergebnisse der Studie übertragbar und helfen sie bei der klinischen Problemlösung meines/-r Patienten/-in?

1. Sind die Ergebnisse der Therapiestudie valide, d.h., sind die methodische Ausführung korrekt und die Ergebnisse frei von Verzerrungen?

Die Richtigkeit der Studienresultate in Bezug auf die untersuchten Umstände wird «*interne Validität*» genannt. Diese sagt aus, dass beobachtete Unterschiede zwischen den Vergleichsgruppen – abgesehen von zufälligen Fehlern – ausschliesslich durch die geprüften Behandlungen zustande kommen. Zur Beurteilung dieser methodischen Qualität sind folgende Fragen zu potentiellen Quellen von Verzerrungen der Studienresultate hilfreich (Tab. 1):

Tabelle 1.
Kriterien zur Beurteilung einer Studie zu einer therapeutischen Massnahme: Validität der Methodik.

Sind die Ergebnisse dieser Therapiestudie valide, d.h., sind die methodische Ausführung korrekt und die Ergebnisse frei von Verzerrungen?
Geschah die Zuteilung der Patienten in die verschiedenen Gruppen nach dem Zufallsprinzip, d.h., handelt es ich um eine randomisierte Studie? Wurde das Randomisierungsschema geheimgehalten?
Waren die Gruppen zu Beginn der Studie vergleichbar?
War das Studiendesign einfach- oder doppelblind?
Wurden die Patienten mit Ausnahme des getesteten Medikaments gleich behandelt?
Wurden alle in die Studie aufgenommenen Patienten in die Schlussevaluation (statistische Analyse) miteinbezogen und in derjenigen Gruppe, in welche sie zu Beginn der Studie eingeteilt wurden, ausgewertet (Intention-to-treat-Analyse)?

Geschah die Gruppenzuteilung der Patienten nach dem Zufallsprinzip, d.h., handelt es sich um eine randomisierte Studie? Wurde das Randomisierungsschema geheimgehalten?

In einer randomisierten Studie besteht für alle Patienten die gleiche Chance, das Testmedikament oder das Vergleichspräparat (Plazebo oder Standardbehandlung) zu erhalten. Dies setzt einen *zufälligen*, *nicht vorhersagbaren* Zuteilungsplan voraus. Die Eleganz der Randomisation besteht darin, dass sie bekannte und *unbekannte* Einflussgrössen des Studienresultats gleichmässig zwischen Studien- und Kontrollgruppe verteilt. Studien ohne Randomisierung können ausgeprägtere Therapieresultate zeigen. Beispielsweise ist der Nutzen der postmenopausalen Östrogensubstitution zur *Primärprävention* des Myokardinfarkts durch Langzeit-Beobachtungsstudien belegt. Diese Ergebnisse können jedoch beschönigt sein, da sich Frauen, die Östrogene einnehmen, aufgrund bestimmter Charakteristika (besseres Gesundheitsbewusstsein, höhere Sozialschicht usw.) wahrscheinlich von Frauen ohne Östrogensubstitution unterscheiden. Dieser wichtige Verzerrungseffekt wird als *Selektions-Bias* bezeichnet. Der Begriff *Bias* meint Vorgänge, welche in einer Studie zu *systematischen* Fehlern führen – im Gegensatz zu *zufälligen* Fehlern. Die Tatsache, dass zwei kürzlich veröffentlichte randomisierte Studien zur Östrogensubstitution in der *Sekundärprävention* des Myokardinfarkts bei postmenopausalen Frauen keinen Nutzen zeigten, legt wichtige Verzerrungseffekte in Beobachtungsstudien der *Primärprävention* des Herzinfarkts mit Östrogenen nahe [4, 5].

Eine methodisch korrekte Randomisierung setzt voraus, dass den Studienmitarbeitern während der Patientenselektion das Randomisierungsschema nicht bekannt ist (*concealment of allocation*). Anderenfalls besteht die Gefahr der bewussten oder unbewussten Selektion: Die Kenntnis der zukünftigen Gruppenzuteilung kann Studienärzte beim Entscheid, Patienten in die Studie aufzunehmen, beeinflussen. Der Studieneintritt bestimmter Patienten wird z.B. so lange verzögert, bis gemäss Randomisationsplan die «geeignete» Behandlung erhältlich wird. Eine Gruppenzuteilung nach vorhersagbaren Mechanismen wie Eintrittstag, Patientennummer oder Geburtsdatum ist deshalb unzulässig. Nur eine Randomisierung, bei welcher für Studienärzte keine Möglichkeit des Zugriffs auf die Patientenzuordnung besteht, ist verlässlich. Am verlässlichsten ist eine sogenannte *zentrale Randomisierung*, welche durch eine unabhängige Koordinationsstelle durchgeführt wird. Bei unserer Studie handelt es sich um eine multizentrische randomisierte Studie, präzise Angaben, wie die Randomisierung vorgenommen wurde, fehlen jedoch.

Waren die Gruppen zu Beginn der Studie vergleichbar?

Insbesondere in kleineren Studien kann es aufgrund der Randomisierung, d.h. durch einen zufälligen Fehler, zu Ungleichheiten bezüglich prognostisch wichtiger Charakteristika kommen, welche die Resultate der Studie beeinflussen können. Frühere Studien und klinische Erfahrung geben Aufschluss darüber, welche Charakteristika Einfluss auf die Prognose eines Patienten haben. Im vorliegenden Fall sind es zumindest Geschlecht, Alter sowie Schweregrad der Depression.
Die Charakteristika Geschlecht, Alter und Schwere der Depression der beiden gebildeten Gruppen von 157 bzw. 167 Patienten waren weitgehend vergleichbar.

War das Studiendesign einfach- oder doppelblind? und
Wurden die Patienten mit Ausnahme des getesteten Medikaments gleich behandelt?

Die ärztliche Erfolgserwartung kann die Beurteilung klinischer Endpunkte verfälschen. Deshalb ist es wichtig, dass die klinischen Endpunkte von Untersuchern bestimmt werden, denen der Behandlungsstatus nicht bekannt ist. Systematische Untersuchungen zeigen, dass Studien, bei welchen die Analyse von Endpunkten nicht verblindet wurde, einen höheren Behandlungsnutzen aufweisen [6, 7, 8]. Auch gefährden Begleitinterventionen, welche eine Vergleichsgruppe bevorzugen, die korrekte Einschätzung des Behandlungseffektes. Diese Gründe sprechen für eine Verblindung der behandelnden Ärzte.
Die gleichzeitige Verblindung der Patienten ist ebenso bedeutsam. Diese vereitelt das Bemühen von Patienten, z.B. in der Beurteilung des Behandlungserfolgs, den Erwartungen des Arztes zu entsprechen. Ist eine Patienten- und Therapeutenverblindung unmöglich (z.B. bei der Evaluation von chirurgischen Therapien), sollten zumindest die Ärzte verblindet sein, welche die klinischen Endpunkte beurteilen.
Das Studiendesign in unserem Beispiel war doppelblind, die Patienten beider Interventionsgruppen wurden ausser den geprüften Medikamenten gleich behandelt. Auch war die in dieser Studie besonders bedeutsame Verblindung bei der Beurteilung des Therapieeffekts am Studienende gegeben («*Blinded outcome assessment*»).

Wurden alle in die Studie aufgenommenen Patienten in die Schlussanalyse miteinbezogen und in derjenigen Gruppe, in welche sie zu Beginn der Studie eingeteilt wurden, ausgewertet (Intention-to-treat-Analyse)?

Falls Patienten, welche nach Randomisierung von der Studie ausgeschlossen werden oder «verloren gehen», in der Schlussevaluation nicht berücksichtigt werden, kann das wahre Behandlungsergebnis verfälscht werden. Eine Verschlechterung der Grundkrankheit, Nebenwirkungen, aber auch ein abgeklungener Leidensdruck können einen Studienabbruch begünstigen [1, 2]. Patienten mit ungenügender Adherance (d.h. unregelmässiger Medikamenteneinnahme) haben häufig eine schlechtere Prognose, unabhängig vom Einfluss der verabreichten Medikamente. Deren Ausschluss kann die Aussagekraft einer Studie durch die Selektion von Patienten mit geringeren Risiken verfälschen. Manchmal allerdings müssen Patienten, für die der Endpunkt nicht erhoben werden konnte, ausgeschlossen werden. Als Faustregel gilt, dass bei Studienabbruchs-Raten über 20% Studienergebnisse mit Vorsicht interpretiert werden müssen [10]. Wenn immer möglich sollten bei Patienten, welche vorzeitig aus einer Studie ausscheiden, die letzten beobachteten Befunde in das Endresultat einfliessen (*last observation carried forward*). Um das Prinzip der Randomisation nicht zu brechen, sollten alle Patienten gemäss ihrer ursprünglichen Zuordnung, d.h. nach dem Intention-to-treat-Prinzip analysiert werden. Falls dies nicht gemacht wird, wird es z.B. in einer Hausgeburtsstudie zur Beschönigung der Ergebnisse der Hausgeburtsgruppe kommen, wenn komplikationsbehaftete Schwangerschaften aus der Hausgeburtsgruppe, die eine Hospitalisation erfordern, nachträglich in der Spitalgruppe analysiert werden.
In der vorliegenden Studie betrug bei 324 Patienten die Drop-out-Rate gesamthaft 15%, wobei davon 19% auf die Imipramin-Gruppe entfielen. Von allen 47 Patienten flossen die letzten beobachteten Befunde in das Endresultat ein, alle Patienten wurden in derjenigen Gruppe analysiert, zu welcher sie bei der Randomisation zugeordnet wurden, es erfolgte also eine Intention-to-treat-Analyse.

2. Sind die Ergebnisse dieser Studie klinisch bedeutsam?

Die zweite Hauptfrage zielt auf die Abschätzung des konkreten Nutzens einer therapeutischen Intervention und beinhaltet die folgenden Punkte:

Welches sind die Endpunkte?
Sind diese, falls valide, klinisch relevant?
Wurden alle klinisch relevanten Endpunkte gemessen?

Beispiele für klinisch relevante, sich am Patientennutzen orientierende Parameter sind Mortalität, Hospitalisation, Funktionsfähigkeit

(z.B. Gehfähigkeit usw.) oder Lebensqualität. Vielfach wird die Wirksamkeit einer Intervention mit Ersatz-Endpunkten, sogenannten Surrogatmarkern, überprüft. Beispielsweise wird die Wirksamkeit von Medikamenten zur Prävention der postmenopausalen Osteoporose mittels Knochendichtemessungen und nicht mit dem Nachweis der Frakturreduktion untersucht. Studien mit Surrogatmarker-Endpunkten können jedoch zu falschen Schlussfolgerungen führen. Bei der Osteoporosetherapie ist z.B. der Zusammenhang zwischen Modifikation der Knochendichte und Beeinflussung der Frakturhäufigkeit in verschiedenen Medikamentenklassen unterschiedlich. Obwohl Fluoride im Vergleich zu Biphosphonaten einen vergleichbaren Effekt auf die Knochendichte bewirkten, erlitten Patientinnen in Langzeitstudien unter Fluoriden eher häufiger Frakturen als unter Plazebo [9].
Als Zielgrössen wurden die «Hamilton Depression Scale», Skalen zur Globalbeurteilung und eine Verträglichkeitsskala herangezogen. Diese validierten Erhebungsinstrumente messen klinisch relevante Befindlichkeitsqualitäten und das Ausmass der Nebenwirkungen.

Wie wurden diese Zielgrössen gemessen?

Die Mortalität ist ein einfach zu bestimmender Parameter. Klinische Endpunkte von Morbiditätsdaten oder des Befindens wie Lebensqualität und Beschwerden sind schwieriger objektivierbar, lassen sich mit validierten Messinstrumenten jedoch ebenfalls quantifizieren. Der Leser sollte jede Studie bezüglich der klinischen Relevanz der gemessenen Endpunkte beurteilen und sich fragen, ob die dafür verwendete Messmethoden geeignet waren [13].
In unserer Studie sind die gemessenen Endpunkte für die gewählte Fragestellung relevant, die verwendeten Messmethoden waren geeignet [11, 12].

Wie gross ist der quantitative Nutzen der Untersuchungsergebnisse?

Endpunkte (z.B. Mortalität) werden oft in Prozentzahlen angegeben. Daraus kann die *absolute Risikoreduktion (ARR)* errechnet werden. Das am häufigsten verwendete Mass zur Beurteilung eines Behandlungsnutzens ist *die relative Risikoreduktion* (*RRR*). Ein dritter Parameter, der das Ausmass des Therapieerfolges anschaulich wiedergibt, ist die sogenannte «*Number Needed to Treat» (NNT)* [10]. Dieser Wert besagt, wieviel Patienten während einer definierten Zeitspanne (z.B. 1 Jahr) behandelt werden müssen, um ein Ereignis zu verhindern.

Die *ARR* entspricht der Differenz der Ereignisrate pro Zeiteinheit (z.B. Mortalität) zwischen Kontroll- *(CER: Control Event Rate)* und Testgruppe *(EER: Experimental Event Rate)*.
ARR(%) = CER(%)–EER(%) pro Zeiteinheit (z.B. 1 Jahr)

Die *RRR* ist die proportionale Abnahme der Ereignisraten zwischen den beiden Gruppen und wird folgenderweise errechnet:
RRR(%) = (CER–EER)/CER(%)
= ARR/CER(%) pro Zeiteinheit (z.B. 1 Jahr)

Die *NNT* entspricht der Anzahl Patienten, die mit Testsubstanz behandelt werden müssen, um pro Zeiteinheit ein Ereignis zu verhindern. Sie entspricht dem reziproken Wert der *ARR*.
NNT = 1/ARR [oder 100%/ARR(%)] pro Zeiteinheit (z.B. 1 Jahr)

Die Angabe von Therapieeffekten mittels absoluter Risikoreduktion oder der NNT gibt uns im Gegensatz zur relativen Risikoreduktion eine entscheidende Information über das *basale Erkrankungsrisiko*. Bei einer sehr tiefen Ereignisrate von beispielsweise 10/1000 pro Jahr bedeutet eine 50% relative Risikoreduktion eine nur geringfügige Senkung des Erkrankungsrisikos auf 5/1000 (ARR von 0,5%, NNT von 200). Bei einer Ereignisrate von 100/1000 pro Jahr bedeutet die gleiche 50%-Risikoreduktion eine Senkung des Erkrankungsrisikos auf 50/1000 (ARR von 5%, NNT 20). Das Beispiel zeigt, warum Angaben in *ARR* bzw. der davon abgeleiteten *NNT* wesentlich informativer sind.
Das *95%-Konfidenzintervall* (95%-CI*)* veranschaulicht die Präzision des Studienresultates. Das Konfidenzintervall ist eine Funktion der Studiengrösse und somit der statistischen Kraft und der Effektgrösse eines erwarteten Wirkungsunterschiedes. Es kann als Wertebereich interpretiert werden, in welchem der «wahre» Wert mit der Wahrscheinlichkeit von 95% liegt.
Häufig findet als Parameter der «statistischen Signifikanz» der «p-Wert» Verwendung. Was bedeutet er?
Bei einer statistischen Auswertung einer randomisierten Studie mit Vergleich zweier Therapien A und B wird von einer Nullhypothese ausgegangen, die besagt, dass Medikament A gleich wirksam ist wie Medikament B. Nehmen wir an, die mit Medikament A behandelten Patienten haben gegenüber den mit Medikament B behandelten ein um 25% vermindertes Risiko zu erkranken. Mit der Testung der «statistischen Signifikanz» überprüfen wir, ob diese 25%-Risikoabnahme einem «wahren» biologischen Effekt zu Grunde liegt oder aber durch Zufall entstanden sein könnte. Hierzu definieren wir einen (willkürlichen) Grenzwert. Bei «p = 0,05» nehmen wir eine Wahrscheinlichkeit von 5% an respektive lassen eine Chance von 1 in 20 zu, dass der gefundene Unterschied

zwischen Therapie A und B rein zufällig zustande kam. Anders formuliert, mit einem p-Wert von 0,05 lassen wir eine Wahrscheinlichkeit von 5% zu, dass wir die Nullhypothese (welche keinen Effekt postuliert) fälschlicherweise ablehnen (Typ-I-Fehler). Die Angabe des p-Wertes gibt somit nur eine Aussage darüber, bei welcher Irrtumswahrscheinlichkeit die Nullhypothese verworfen werden kann. Demgegenüber gibt das 95%-Konfidenzintervall dem Kliniker eine wichtige Zusatzinformation über den möglichen Streubereich eines gefundenen Wirkungseffektes. Angenommen das 95%-Konfidenzintervall der oben erwähnten 25%igen relativen Risikoreduktion reiche von 2% bis 50% bei einem p-Wert von 0,04. Obwohl das Ergebnis statistisch «signifikant» ist, muss im pessimistischen Fall von praktisch keinem Wirkungseffekt bei einer relativen Risikoreduktion von nur 2% ausgegangen werden.

Die Autoren der vorliegenden Studie fanden in der «Hamilton Skala» unter Hypericum und Imipramin eine Abnahme des Depressions-Scores von ursprünglich rund 22 Punkten um 12,0 respektive 12,75 Punkte auf rund 10 Punkte. Dies entspricht einer Abnahme von einer mittelschweren Depression zu einer leichten Depression. Die Differenz von –0,75 Punkten (95% CI –1,90 bis +0,40 Punkte) ist statistisch nicht signifikant und klinisch nicht relevant. Ähnlich wie bei der ARR können aus dem direkten Vergleich der Ergebnisse Schlüsse gezogen werden. Dies setzt allerdings voraus, dass für Kliniker nützliche Vergleichsmöglichkeiten vorliegen, ob ein gefundener Unterschied in einem Score klinisch relevant ist [13]. Zudem setzt deren Gewichtung Erfahrung mit diesen Messskalen voraus.

Hyperikum zeigt im Vergleich zu Imipramin bezüglich Nebenwirkungen einen statistisch signifikanten günstigeren Effekt (Relative Risikoreduktion für den Abbruch der Medikation aufgrund von Nebenwirkungen = 81%). Auf 8 mit Hypericum behandelte Patienten konnten gegenüber Imipramin bei einem Patienten relevante Nebenwirkungen verhütet werden, welche zum Therapieabbruch geführt hätten (NNT = 8). In Tabelle 3 sind die Berechnungen zur Unterschiedlichkeit der Nebenwirkungen aufgeführt. Eine Demonstrationsversion des Programmes «CATmaker» (critical appraised topics), welches die für diese Berechnungen notwendigen Zahleneingaben ausführt, ist unter folgender Website erhältlich: http://cebm.jr2.ox.ac.uk.

Wurde die Testsubstanz mit einer Standardbehandlung oder mit Plazebo verglichen?

Bei der Beschreibung einer neuen Therapie interessiert der Vergleich mit der bisherigen Standardbehandlung (in adäquater Dosierung), wenn eine solche zur Verfügung steht. Vergleiche mit Plazebo zeigen meist eindrücklichere Wirkungen, so dass in Studien manchmal neue Behandlungen gegen Plazebo geprüft werden, obwohl wirksame Standardtherapien zur Verfügung stehen. Ergebnisse von solchen Studien weisen eine geringere Relevanz auf und werfen auch grundsätzliche ethische Fragen zur deren Vertretbarkeit auf [14].

Die experimentelle Therapie wurde mit einer Standardtherapie in adäquater Dosierung verglichen.

Tabelle 2.
Kriterien zur Beurteilung einer Studie zu einer therapeutischen Intervention: Relevanz der Studienergebnisse und Umsetzung der Ergebnisse auf Ihre eigenen Patienten (externe Validität).

Sind die Ergebnisse dieser Studie klinisch bedeutsam?
Kriterien zur Beurteilung des Behandlungsnutzens einer Studie zu einer therapeutischen Intervention:
Welches sind die Zielgrössen? Sind diese, falls valide, klinisch relevant?
Wie wurden diese Zielgrössen gemessen?
Wie gross ist der potentielle quantitative Nutzen?
Wurde die Testsubstanz mit einer Standardbehandlung oder mit Plazebo verglichen?
Sind die Ergebnisse dieser Studie übertragbar und helfen sie bei der klinischen Problemlösung meines/-r Patienten/-in?
Sind die Resultate auf das Problem oder die Frage meines/-r Patienten/-in übertragbar?
Entsprechen die Merkmale meines/-r Patienten/-in den Patienten der Studie?
Kann mein/e Patient/in von der in der Studie beschriebenen Endpunkt-Verbesserung profitieren?
Überwiegt der zu erwartende Therapieeffekt bei meinem/-r Patienten/-in die Therapierisiken und wie hoch wäre der Preis eines Therapieverzichts?

3. Sind die Ergebnisse dieser Studie übertragbar und helfen sie bei der klinischen Problemlösung meines/-r Patienten/-in?

Die Überprüfung der praktische Umsetzbarkeit von Studienergebnissen zum Nutzen unserer Patienten/-innen im klinischen Alltag ist sehr wichtig. Fragen zu Verallgemeinerungsfähigkeit von Studienresultaten in der Praxis werden mit dem Begriff der «*externen Validität*» umschrieben [1, 2, 6, 7]. Die Beantwortung folgender Fragen ist bei der Beurteilung der äusseren Validität hilfreich (Tabelle 2):

Sind die Resultate auf Ihre Patientin übertragbar?
Entsprechen die Merkmale Ihrer Patientin den Patienten der Studie?

Kann Ihre Patientin von der in der Studie beschriebenen Endpunkt-Verbesserung profitieren?
Falls die Studienresultate auf Ihre Patientin übertragbar sind, überwiegt der Therapieeffekt die Therapierisiken und wie hoch wäre der Preis eines Therapieverzichts?

In dieser Studie wurden erwachsene ambulante Patienten aus Grundversorgerpraxen in Deutschland mit leicht bis mässig ausgeprägter Depression untersucht. Dies entspricht dem Profil Ihrer Patientin. Die untersuchte Therapiedauer von sechs Wochen ist allerdings zu kurz. Eine längere Beobachtungsperiode, insbesondere bezüglich Nebenwirkungen und Interaktionen (Hyperikum aktiviert das Zytochrom P450), wäre aufschlussreicher. Bezüglich Verträglichkeit und zu erwartender Therapieeffekte bezüglich Ihrer Patientin sehen Sie gegenüber den Studienpatienten eine vergleichbare Ausgangslage. In der ambulanten Grundversorgung spielen bei leichten bis mittelschweren Depressionen aufgrund des günstigeren Nebenwirkungsprofils und einfacherem Einnahmemodus SSRI jedoch eine zunehmend grössere Rolle als Trizyklika. Deswegen wäre ein Vergleich von Hyperikum / SSRI informativer.

Tabelle 3.
Relatives und absolutes Risiko einer zum Therapieabbruch führenden Nebenwirkung von Hypericum im Vergleich zu Imipramin [3].

	Nebenwirkungen mit Abbruch der Medikamenteneinnahme: Ja	Nein	
Hypericum	4 (a)	153 (b)	157 (a+b)
Imipramin	26 (c)	141 (d)	167 (c+d)
	30 (a+c)	294 (b+d)	Alle Patienten 324 (a+b+c+d)
	Formel	Berechnung	(95% CI)
Inzidenz Interventionsgruppe	a/(a+b)	4/157×100% = 3%	
Inzidenz Kontrollgruppe	c/(c+d)	26/167×100% = 15.6%	
Relatives Risiko	a/(a+b) : c/(c+d)	(0.03/0.156)×100% = 19%	(6.8–53.2%)
Relative Risikoreduktion	$\frac{c/(c+d) - a/(a+b)}{c/(c+d)}$	$\frac{0.156 - 0.03 \times 100\%}{0.156}$ = 81%	(42–100%)
Absolute Risikoreduktion	[c/(c+d) – a/(a+b)]×100%	15.6% – 3% = 12.6%	(6.5–18.7%)
NNT	1/[c/(c+d) – a/(a+b)]	1/0.126 = 8	(5–15)

Schlussfolgerung

Trotz der genannten Schwachpunkte bezüglich interner und externer Validität *(keine genauen Angaben zur Randomisation, kurze Studiendauer, Problematik des Standardmedikaments)* darf aus dieser Studie geschlossen werden, dass bei leichten bis mässig ausgeprägten Depressionen adäquate Dosen von Hyperikum bzw. Imipramin mit grosser Wahrscheinlichkeit eine ebenbürtige Wirkung aufweisen. Diese Beurteilung von Johanniskraut entspricht weitgehend Erkenntnissen bisheriger Studien und einer Meta-Analyse [15, 16]. Aufgrund dieser Evidenz gelangen Sie zum Entscheid, dem Vorschlag Ihrer Patientin mit mässig ausgeprägter Depression zu folgen und das Phytotherapeutikum Hyperikum einzusetzen. Allerdings warten Sie gespannt auf eine grössere randomisierte kontrollierte Studie, bei der bei einem ähnlichen Patientenspektrum über einen längeren Zeitraum die Therapieeffekte von Hyperikum mit einem SSRI verglichen werden.

Literatur

1 Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1993;270:2598–601.
2 Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 1994;271:59–63.
3 Helmut Woelk. Comparison of St John's wort and imipramine for treating depression: randomised controlled trial. BMJ 2000;321: 536–9 (http://bmj.com/cgi/content/full/321/7260/536).
4 Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomised trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research JAMA 1998; 280:605–13.
5 Herrington DM, Reboussin DM, Brosnihan KB, Sharp PC, Shumaker SA, Snyder TE, et al. Effects of estrogen replacement on the progression of coronary-artery atherosclerosis. N Engl J Med 2000; 343: 522–9.
6 Schulz KF, Chalmers I, Hayes RJ, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
7 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054–60.
8 Moher D, Ba'Pham, Jones A, Cook DJ, Jahad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analysis? Lancet 1998;352:609–13.

9 Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. JAMA 1999;282:771–8.
10 Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine. How to practice and teach EBM. New York, Edinburgh, London: Churchill Livingstone, 2000.
11 Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62.
12 National Institute of Mental Health. 028 CGI clinical global impressions. In: Guy W, ed. EDC-EU assessment for psychopharmacology. Rev. Ed. Rockville, MD: National Institute of Mental Health, 1976: 217–22.
13 Guyatt GH, Naylor CD, Cook DJ. Users' guides to the medical literature. VII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. JAMA 1997;277:1232–7.
14 Rothman KJ, Michels KB. The Continuing Unethical Use of Placebo Controls. N Engl J Med 1994; 331: 394–8.
15 Begg C, Cho M, Eastwood S, Horton R, Moher 0, 0lkin I, et al. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. JAMA 1996; 276:637–9.
16 Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D. St John's Wort for depression – an overview and meta-analysis of randomised clinical trials. BMJ 1996;313:253–8.