3 Wirksamkeit und Nutzwert diagnostischer Test und ihre Abhängigkeit von der Einhaltung methodischer Standards

3.1. Forschungsergebnisse

3.1.1. Kernaussagen

- Die seit vielen Jahren klinisch etablierte Ultraschalluntersuchung des Abdomens zur Diagnose des stumpfen Bauchtraumas hat eine zu geringe Sensitivität, um weiterhin als Standardmethode propagiert zu werden.
- Trotz einer Verbesserung der Sensitivität durch Ausschöpfung der technischen Möglichkeiten (differenzierter Einsatz verschiedener Schallköpfe) kann bei einem negativen Befund nicht auf weitere bildgebende Diagnostik verzichtet werden.
- Neben einer erheblichen Publikationsverzerrung (d.h., dem Fehlen negativer Studien in der medizinischen Literatur) besteht eine starke Assoziation zwischen der methodischen Qualität der publizierten Arbeiten und der berichteten diagnostischen Genauigkeit.
- Ultraschall-basierte Algorithmen führen im Vergleich zu nicht-Ultraschall-basierten Algorithmen nicht zu einer Verbesserung der Ergebnisse in patientenzentrierten Endpunkten.

3.1.2. Publikationen

- **Stengel D**, Bauwens K, Rademacher, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. *Radiology* 2005 236:102-111.
- **Stengel D**, Bauwens K, Porzsolt F, Rademacher G, Mutze S, Ekkernkamp A. Emergency ultrasound-based algorithms for diagnosing blunt abdominal trauma. *Cochrane Database Syst Rev* 2005 Apr 18;(2):CD004446.
- **Stengel D**, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F. A likelihood ratio approach to meta-analysis of diagnostic studies. *J Med Screen* 2003;10:47-51.
- **Stengel D**, Bauwens K, Sehouli J, Nantke J, Ekkernkamp A. Discriminatory power of 3.5 MHz convex and 7.5 MHz linear ultrasound probes for the imaging of traumatic splenic lesions: a feasibility study. *J Trauma* 2001;51:37-43.
- **Stengel D**, Bauwens K, Sehouli J, Porzsolt F, Rademacher G, Mutze S, Ekkernkamp A. Systematic review and meta-analysis of emergency ultrasonography for blunt abdominal trauma. *Br J Surg* 2001;88:901-912.

3.2. Verzerrte Wahrnehmung des Nutzens diagnostischer Tests

3.2.1. Die Notwendigkeit für Systematische Reviews und Meta-Analysen

Biologisch plausible, wenig invasive, einfach zu handhabende und generell verfügbare Methoden stellen den Idealfall eines diagnostischen Tests dar. Erfüllt eine Methode diese Kriterien, wird ihre Effektivität selten hinterfragt oder durch veraltete bzw. selektiv zitierte Daten gestützt.

Die Befürworter der Ultraschalluntersuchung argumentieren, dass

- wenn die Methode keine relevante Information erbringt, sie doch zumindest nicht schadet,
- die Bewertung ihres Ergebnisses lediglich von der Erfahrung des Untersuchers abhängt,
- die generelle Anwendung tatsächlicher Referenzstandards zu umständlich, logistisch unmöglich oder zu teuer ist.

Diese Argumente können und dürfen in der modernen Medizin keinen Bestand mehr haben. Ein für Arzt, Patient und Gesellschaft überflüssiger Test muss aus dem Katalog angebotener Gesundheitsleistungen gestrichen werden. Spezialisierung und Erfahrung ist ein wesentlicher Bestandteil ärztlicher Tätigkeit (insbesondere in der operativen bzw. interventionellen Medizin und der Bildgebung). Kann jedoch ein Testergebnis ausschließlich durch erfahrene Untersucher bewertet werden, muss an der generellen Anwendbarkeit einer Methode gezweifelt werden. Ob zuletzt nicht die primär definitive Sicherung einer Diagnose bzw. ihr Ausschluss in einem Untersuchungsschritt für alle Beteiligten erheblich weniger belastend, zügiger und kostengünstiger ist, sollte dringend hinterfragt werden.

Für den Nachweis der Effektivität einer Intervention müssen alle verfügbaren Daten zusammengestellt und mit klinischer und methodischer Expertise bewertet werden. Ohne das Instrument der systematischen Übersichtsarbeit und Meta-Analyse können Gesundheitsleistungen nicht mehr bewertet werden.

Im Gegensatz zu Meta-Analysen von randomisierten Interventionsstudien, für die nicht zuletzt dank der Cochrane Collaboration strukturierte Anleitungen zu ihrer Erstellung und Bewertung etabliert wurden, ergeben sich für diagnostische Meta-Analysen folgende Probleme:

- keine einheitliche Suchstrategie in medizinischen Datenbanken bzw. keine präzise Indexierung von Diagnosestudien,
- keine allgemein akzeptierten Instrumente zur Bewertung der methodischen Qualität identifizierter Studien,
- klinisch schwierig zu interpretierende gemeinsame Effektschätzer.

Meta-Analysen sollten neben einer transparenten (zumeist grafischen) Darstellung der Ergebnisse einzelner Studien ein mitteilbares Gesamtergebnis liefern, das Anwendern und Kostenträgern die Entscheidung für oder gegen ein Verfahren erleichtert. Aus therapeutischen Meta-Analysen ist dies unschwer abzuleiten- eine Intervention wirkt im Gegensatz zu einer Kontrolle besser oder schlechter, bzw. existiert kein Anhalt für eine Differenz zwischen beiden Verfahren.
Die klinisch noch wenig thematisierte methodisch einwandfreie Aufarbeitung der besten verfügbaren Evidenz zu einem diagnostischen Test sollte anhand eines praktischen Beispiels verfolgt werden.

Aus der eigenen langjährigen Erfahrung mit der Ultraschalluntersuchung bei Schwerstverletzten und Expertendiskussionen wurden Fragen formuliert, die mit meta-analytischen Methoden beantwortet werden sollten. Ein weiteres Ziel war es, nicht nur eine zweifelhafte bzw. mangelnde diagnostische Genauigkeit offen zu legen, sondern auch technische Modifikationen zu untersuchen, mit deren Hilfe die Effektivität der Sonografie möglicherweise verbessert werden könnte. Die Ergebnisse sollten in einer sowohl methodisch einwandfreien als auch klinisch akzeptablen Form dargestellt werden.

3.2.2. Qualitative und quantitative Bewertung der Ultraschalluntersuchung zum Nachweis und Ausschluss abdomineller Verletzungen

3.2.2.1. Hintergrund

Die Ultraschalluntersuchung des Abdomens im Schockraum in drei oder vier Standardebenen (Focused Abdominal Sonography for Trauma, FAST) zum Nachweis von freier intraabdomineller Flüssigkeit (d.h. Blut) oder Organverletzungen hat sich seit ihrer Einführung in Europa in den frühen 80er Jahren als Standardprozedur im diagnostischen Algorithmus schwerstverletzter Patienten etabliert und ist Bestandteil nationaler und internationaler Leitlinien.[31,32] In

[31] Leitlinien-Kommission der Deutschen Gesellschaft für Unfallchirurgie e.V. Polytrauma. Leitlinie für die Unfallchirurgische Diagnostik und Therapie. *Unfallchirurg* 2001;104:902–912.

den Europäischen Ländern hat das FAST-Protokoll den früher üblichen invasiven Nachweis von Blut in der freien Bauchhöhle (Hämoperitoneum) durch die diagnostische Peritoneallavage ersetzt.

Im Gegensatz zur unter elektiven Bedingungen durchgeführten Ultraschalluntersuchung zielt das FAST-Regime nicht auf die detaillierte Beschreibung des Verletzungsmusters ab, sondern soll lediglich diejenigen Patienten mit dringlich operationspflichtigen Befunden identifizieren. Blut wird hierbei als Surrogat für eine traumatische Organschädigung angesehen.
Die Methode kommt bei Organverletzungen ohne Blutaustritt in die freie Bauchhöhle jedoch an ihre Grenzen, insbesondere dann, wenn der Untersucher ungeübt und die Umgebungsbedingungen (Lichteinflüsse, Zeitfaktor) ungünstig sind. In der internationalen Literatur reicht die Prävalenz dieser Form von Bauchverletzungen von 4,9 bis 37,1%. Hinzu kommt, dass eine beim wachen Verletzten durch peritoneale Reizung vermittelte Abwehrspannung der Bauchdecken durch die pharmakologische Muskelrelaxierung und Analgesie bei Beatmeten aufgehoben wird.
Abdominelle Organschäden nach stumpfem Bauchtrauma zählten bis vor zehn Jahren zu den am häufigsten übersehenen Verletzungen.[33] Ob die Einführung der Ultraschalluntersuchung dieses Problem global korrigieren konnte, ist derzeit noch ungewiss.

Die Befürworter einer sofortigen CT-Untersuchung Schwerstverletzter weisen auf die unzureichende diagnostische Genauigkeit der Ultraschalluntersuchung hin. Ihre Protagonisten argumentieren, dass die Sonografie in Zusammenschau mit klinischen Befunden und der Hämodynamik bewertet werden muss.
Eine Kreislaufinstabilität in Verbindung mit abdominellen Gurtmarken wird an einem positiven Sonogramm kaum Zweifel aufkommen lassen.
Wenn auch hier nur für stumpfe Leberverletzungen gezeigt, gibt es jedoch keine plausiblen Gründe, warum der *fehlende klinische Nachweis* sog. Indikatorverletzungen die Wahrscheinlichkeit für traumatische Schäden von Milz, Pankreas, Dünndarm usw. erniedrigen sollte. Dies gilt umso mehr für den vital bedrohten polytraumatisierten Patienten.
Welche Testcharakteristik der Ultraschalluntersuchung ist in der Notfallsituation relevant? Grundsätzlich muss zwischen

---

[32] American College of Emergency Physicians. Clinical policy for the initial approach to patients presenting with acute blunt trauma. *Ann Emerg Med* 1998; 31: 422–454.

[33] Hodgson NF, Stewart TC, Girotti MJ. Autopsies and death certification in deaths due to blunt trauma: what are we missing? *Can J Surg* 2000;43:130-136.

1) der *Bestätigung* eines klinisch vermuteten Befundes durch einen sehr spezifischen Test mit geringer Rate falsch-positiver Befunde und dem
2) sicheren *Ausschluss* einer Diagnose bei klinisch unauffälligem Befund durch einen sehr sensitiven Test mit geringer Rate falsch-negativer Befunde

unterschieden werden.

Der Arzt muss sich vor dem Hintergrund von Behandlungsprioritäten beim Schwerstverletzten auf einen *negativen* Testbefund verlassen können. Es stellt sich somit führend die Frage, wie *sensitiv* die Ultraschalluntersuchung im Vergleich zu etablierten Referenzstandards ist.

Um eine valide Grundlage für die derzeit kontroverse Diskussion zu schaffen, wurde die diagnostische Effektivität der Sonografie in einer systematischen Übersichtsarbeit und Meta-Analyse untersucht.
Ein weiterer Aspekt dieser Untersuchung, die 1998 begonnen und in den Folgejahren kontinuierlich aktualisiert und unter neuen methodischen Gesichtspunkten reanalysiert wurde, stellt die Abhängigkeit der diagnostischen Kenngrößen von der methodischen Qualität der Datenquellen dar.

Ein aus der Therapieforschung bekannter Effekt ist die Überschätzung der Wirksamkeit einer Intervention durch methodisch mangelhafte Studien.[34] Empirisch konnte in wenigen Untersuchungen ein derartiges Phänomen auch in Diagnosestudien beobachtet werden.[35,36] Über die Generierung von Punktschätzern hinaus sollte daher der Einfluss methodischer Standards auf die berichtete Testgenauigkeit beschrieben werden.

3.2.2.2. Methoden

Prospektive Studien aus dem Publikationszeitraum Januar 1957 – November 2003 wurden unabhängig von der Sprache in elektronischen Datenbanken identifiziert. Als Datenquellen dienten die elektronischen Datenbanken Pubmed Medline (Index

---

[34] MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;4(34):1-154.

[35] Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004; 140:189-202.

[36] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-1066.

Medicus seit 1966) einschl. Oldmedline (Index Medicus 1957 – 1966), das Cochrane Central Register of Controlled Trials (CENTRAL), Embase (Excerpta Medica, seit 1984), Scisearch und Cinahl. Die Suchstrategie umfasste zudem eine Handsuche und eine freie Internet-Recherche.
Um eine eventuelle Publikationsverzerrung (Publication Bias), d.h. das Fehlen kleiner Studien mit nicht-signifikantem Ergebnis, zu erkennen, wurde der von Hasselblad und Hedges vorgeschlagene *d*-Wert als globaler Marker für die Diskriminationsfähigkeit eines diagnostischen Tests berechnet.[37] Die inversen Standardfehler wurden gegen die *d*-Werte der Einzelstudien im Sinne eines Funnel-Plots aufgetragen.[38] Die Funnel-Plot Asymmetrie wurde mit der von Egger et al. vorgeschlagenen Erweiterung des Galbraith-Plots geprüft.[39]

Die Bewertung der methodischen Qualität erfolgte unabhängig durch zwei Reviewer [40]anhand von 27 Kriterien, die den kürzlich vorgeschlagenen Standards for Reporting of Diagnostic Accuracy (STARD)[41] und dem Quality Assessment of Diagnostic Accuracy Studies (QUADAS) Instrument entnommen wurden.[42]
Neben methodischen Standards wurde der Einfluss verschiedener Zielkriterien (d.h., der fokussierte Nachweis freier Flüssigkeit oder die sonografische Abbildung der Organverletzung), demografischer und anderer Studiencharakteristika auf die diagnostische Genauigkeit untersucht.
Hierzu wurden Summary Receiver Operating Characteristics (SROC)[43] ermittelt und die gemeinsame Sensitivität und Spezifität mit 95% Konfidenzintervallen (KI) durch Random-Effects Meta-Regression berechnet.

---

37 Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psych Bull* 1995;117:167-178.

38 Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31:88-95.

39 Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-634.

40 Co-Reviewer und –Autor war in jedem Fall Herr Dr. med. Kai Bauwens, Unfallkrankenhaus Berlin

41 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41-44.

42 Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.

43 Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;12:1293-1316.

Faktoren mit in der univariaten Analyse signifikantem ($p<0,25$) Einfluss auf die Effektschätzer wurden in einer schrittweisen Selektionsprozedur in ein multivariates Meta-Regressions-Modell integriert.

3.2.2.3. Ergebnisse

Von 957 identifizierten Studien enthielten nach Durchsicht von Titel oder Abstract 372 Arbeiten potenziell verwertbare Informationen. Berücksichtigt wurden 73 Publikationen von insgesamt 62 Studien mit Einschluss von 18144 Patienten
Die Prüfung der Funnel-Plot-Asymmetrie wies auf das Vorliegen einer Publikations-Verzerrung hin. Die Studien erfüllten im Median 13 (Interquartilspanne [IQR] 11 – 16) der 27 methodischen Kriterien.
Die Sensitivität war stark heterogen (31 – 100%), die Spezifität hingegen über alle Studien konstant hoch. Die Prävalenz von abdominellen Verletzungen lag bei 25,1%.
Die über alle Studien und für alle Endpunkte (freie intraabdominelle Flüssigkeit oder Organverletzung) gepoolte Sensitivität der Ultraschalluntersuchung betrug 78,7% (95% KI 74,7 – 82,7%), die Spezifität 99,0% (95% KI 98,7 – 99,3%). Hieraus berechnete sich eine positive Likelihood Ratio von 79 und eine negative Likelihood Ratio von 0,22. Ein positives Sonogramm erhöht damit substanziell die Vortest-Wahrscheinlichkeit intraabdomineller Verletzungen, ein negatives Sonogramm schließt Verletzungen nicht aus.
Unter Annahme der berichteten Prävalenz von 25% liegt die Nachtest-Wahrscheinlichkeit für intraabdominelle Verletzungen nach positivem Sonogramm bei 96%, nach negativem Sonogramm bei 7%.
Die linke Grafik in Abbildung 8 zeigt die Streuung der diagnostischen Effektmaße und die gemeinsame ROC Kurve.
Der vermutete Unterschied in der diagnostischen Genauigkeit für unterschiedliche Zielkriterien ließ sich nicht bestätigen.
Sensitivität und Spezifität für den Ausschluss bzw. Nachweis eines Hämoperitoneums lagen bei 79,4% und 99,1%, für Organverletzungen bei 76,5% und 98,4%.

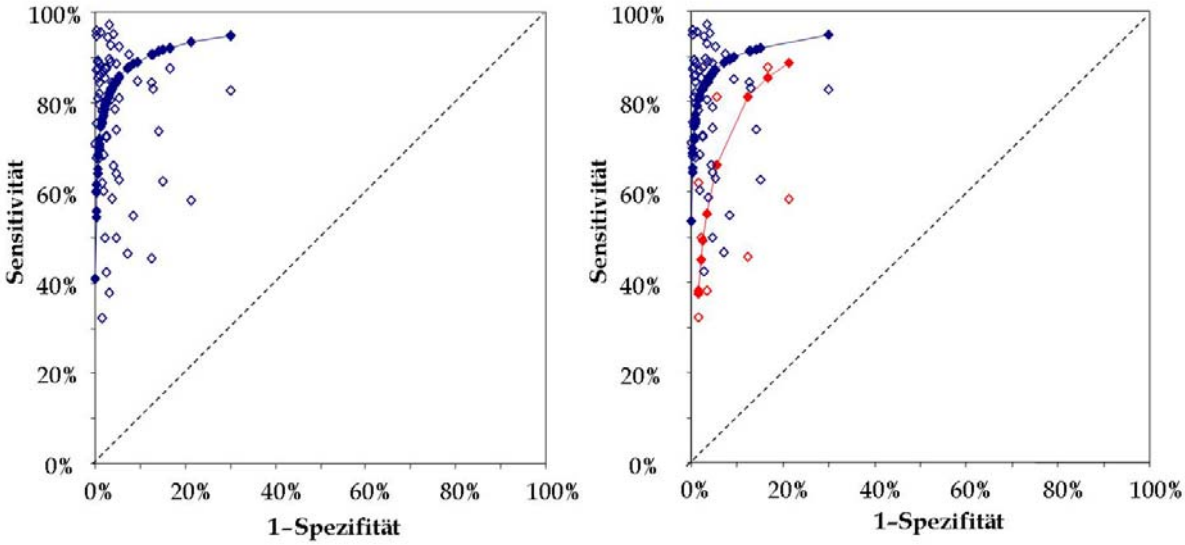

Abbildung 8 Gemeinsame Receiver Operating Characteristics (ROC) der Ultraschalluntersuchung bei vermuteter Abdominalverletzung (im Vergleich zu allen berichteten Referenzstandards). Links: alle Studien. Rechts: Unterschiede in den Testcharakteristika zwischen Erwachsenen (blau) und Kindern (rot).

Die Ultraschalluntersuchung war im kindlichen Patientengut (neun Studien, 937 Patienten) unzuverlässiger als bei Erwachsenen (s. Abbildung 12, rechts). Bei einer Prävalenz von 25,2% (95% KI 17,4 – 32,9%) lagen die gemeinsame Sensitivität und Spezifität bei 61,7% (95% KI 46,9 – 76,4%) und 94,8% (95% KI 90,8 – 98,8%).
Die Nachtest-Wahrscheinlichkeit wird bei einer angenommenen Prävalenz von 25% bei positivem Sonogramm auf 80% erhöht, bei negativem Ultraschall auf 12% reduziert.
Die SROC Kurve zeigte bei Radiologen einen flacheren Verlauf als bei Chirurgen.
Generell tendierten Studien, die bestimmte methodische Aspekte nicht berücksichtigten, zu einer Überschätzung der Sensitivität der Ultraschalluntersuchung. Die Differenz zwischen den Effektschätzern reichte von 9% bis 20% (s. Tabelle 5).
In der Meta-Regression erwiesen sich die Verwendung eines einheitlichen Referenzstandards, die Angabe der behandelten, aber nicht in die Studie eingeschlossenen Patienten und die Angabe von Konfidenzintervallen als unabhängige Einflussgrößen auf die Sensitivität.

Die gemeinsame Sensitivität aus 14 Studien mit weniger als 11 respektierten methodischen Standards (unteres Quantil) betrug 91,1% (95% KI 87,2 – 95,3%). Die

Sensitivität der Ultraschalluntersuchung fiel auf 65,5% (95% KI 56,9 – 74,0%), wenn nur diejenigen 14 Untersuchungen mit mehr als 16 methodischen Standards (oberes Quantil) betrachtet wurden.

Variable	Sensitivität (95% KI), Studien mit fehlendem Standard			Sensitivität (95% KI), Studien mit erfülltem Standard		
Details zum Referenzstandard	83,3%	79,6%	87,1%	66,9%	58,1%	75,7%
einzelner Referenzstandard	82,8%	79,1%	86,5%	65,9%	56,1%	75,7%
Konfidenzintervalle angegeben	81,1%	77,3%	84,9%	61,3%	48,6%	74,1%
Verblindung gegenüber Sonogramm	80,6%	76,5%	84,6%	63,9%	51,4%	76,5%
Angabe von Studien-Ausscheidern	84,5%	79,7%	89,3%	74,2%	68,5%	79,9%
Unabhängige Verifikation	85,4%	80,2%	90,5%	75,2%	69,9%	80,4%
Verblindung	81,4%	77,1%	85,7%	71,1%	62,6%	79,7%
Eindeutig formulierte Hypothese	86,6%	81,6%	91,7%	75,9%	71,1%	80,9%
Nennung behandelter Patienten	81,2%	77,1%	85,3%	71,6%	62,0%	81,3%
Geeigneter Referenzstandard	95,6%	91,9%	99,3%	76,2%	71,0%	81,3%
Detailliertes Ultraschallprotokoll	88,9%	84,7%	93,1%	77,5%	73,2%	81,9%
Verblindung gegenüber Referenz	79,0%	74,4%	83,9%	71,7%	63,8%	79,7%
Rekrutierungsperiode genannt	69,3%	53,0%	85,6%	79,5%	75,5%	83,6%
Konsekutive Kohorte	80,9%	76,5%	85,5%	75,4%	68,1%	82,8%

Tabelle 5 Abhängigkeit der geschätzten Sensitivität der Ultraschalluntersuchung von methodischen Standards.

3.2.2.4. Wissenschaftlicher Erkenntnisbeitrag und Einschränkungen der Interpretation

Die Schockraumdiagnostik muss den Spagat zwischen Kosten und Nutzen, geringer Invasivität bzw. Exposition gegenüber toxischen Substanzen und Röntgenstrahlung und dem *kompromisslosen*, *definitiven Ausschluss* von abdominellen Traumafolgen bewältigen. Anders ausgedrückt darf die potenziell geringere Belastung durch ein diagnostisches Verfahren nicht mit einer niedrigeren Genauigkeit erkauft werden.

Die Ergebnisse dieser Meta-Analyse legen nahe, dass diese Ziele mit der Ultraschalluntersuchung allein nicht erreicht werden können. Ist bereits die Effektivität eines diagnostischen Instruments unzureichend, sollten Wirkungsgrad und Effizienz unbedeutend sein, da in diesem Fall ein besseres diagnostisches Instrument standardmäßig angewendet werden *muss*. Die Resultate stützen die Verfechter einer frühen CT-Diagnostik bei stumpfem Bauch- und Polytrauma.
Entgegen der ursprünglichen Vermutung konnte kein Unterschied zwischen der diagnostischen Effektivität des FAST-Protokolls im engeren Sinn (also dem

fokussierten sonografischen Nachweis freier intraabdomineller Flüssigkeit) und komplexeren Untersuchungstechniken zur direkten Darstellung vor Organverletzungen (Kapselrupturen, Hämatome, Kontusionen) gefunden werden.
Die verfügbaren Studien zur Ultraschalluntersuchung bei vermutetem Abdominaltrauma sind zudem nur von durchschnittlicher methodischer Qualität. Arbeiten mit und ohne Gewährleistung bestimmter Standards zeigen deutliche Differenzen in den diagnostischen Effektschätzern, die führend von der Wahl des Bestätigungstests abhängen. Der Nachweis eines Publication Bias lässt vermuten, dass selbst die adjustierten Punktwerte die diagnostische Genauigkeit noch überinterpretieren.

Im Zeitalter nicht-operativer Therapiestrategien auch ausgedehnter Parenchymschäden von Milz und Leber beim kreislaufstabilen Patienten wird weder der sonografische Nachweis eines Hämoperitoneums, noch einer Organverletzung eine sofortige therapeutische Intervention, sondern vielmehr die vollständige und definitive Beschreibung des Verletzungsmusters nach sich ziehen.
Die Sonografie ist fraglos eine plausible, wenig invasive und kostengünstige Untersuchungsmethode. Ihr Nutzwert wird durch diese Charakteristika allein jedoch nicht belegt. Die verfügbaren Daten sind eindeutig: ein negatives Sonogramm würde beim kreislaufstabilen Patienten nur bei ohnehin sehr niedriger Vortest-Wahrscheinlichkeit den Verzicht auf weiterführende Diagnostik erlauben, im positiven Fall ist die Sonografie beweisend und sollte weiterführende Diagnostik induzieren.

3.3. „Negative“ Meta-Analysen als Motor für klinische Forschung

3.3.1. Motivation statt Nihilismus

Eine weit verbreitete Schlussfolgerung aus Meta-Analysen lautet sinngemäß: „Es existiert keine ausreichende wissenschaftliche Begründung für den Einsatz der Methode. Weitere Studien sind erforderlich, um die Effektivität des Verfahrens zu belegen.“
Wenn auch häufig gerechtfertigt, ist diese Aussage selten hilfreich. Sie führt im klinischen Alltag zu Unsicherheit, Unverständnis und Ablehnung. Der Arzt hat die ethische und moralische Verpflichtung, trotz wissenschaftlicher Unsicherheit zu handeln. Auch sollte das Fehlen adäquater Daten[44] nicht zu Nihilismus verleiten,

[44] Hier gilt der aus der Logik bekannte Satz “Lack of evidence of efficacy does not constitute evidence of lack of efficacy”

sondern vielmehr motivieren, die fehlende Information zu generieren. Es ist ein Fehler der EBM-Bewegung, lediglich auf Mängel hinzuweisen, ohne probate und für den Kliniker verständliche Lösungsansätze vorzuschlagen.[45]

Die in der Meta-Analyse verdeutlichte unzureichende Sensitivität der Ultraschalluntersuchung im Vergleich zum CT im pädiatrischen Krankengut war überraschend und erforderte eine Klärung.
Im klinischen Alltag gelingt die sonografische Untersuchung von Kindern üblicherweise mit hoher Auflösung und Präzision. Gründe hierfür liegen insbesondere in den schlanken Bauchdecken mit geringer Schallauslöschung und wenigen Artefakten bei gleichzeitig günstigerem Breiten-Tiefen-Verhältnis des kindlichen Abdomens. Im Oberbauch-Querschnitt lässt sich häufig bereits eine Panorama-Übersicht mit simultaner Darstellung beider Nieren erzielen.
Die unter elektiven Bedingungen guten technischen Möglichkeiten der Ultraschalluntersuchung werden in der Trauma-Situation durch Lichteinflüsse, Prozeduren wie dem Legen zentraler Venenkatheter und Zeitdruck stark eingeschränkt. Auch wenn die CT fraglos den diagnostischen Referenzstandard zur definitiven Abklärung abdomineller Verletzungen bim schwerstverletzten Kind darstellt, sind sowohl Chirurgen als auch Radiologen verpflichtet, die Strahlenbelastung auf das notwendige Mindestmaß zu reduzieren und die Aussagekraft der minimal belastenden Ultraschalluntersuchung zu verbessern.
Neben der Optimierung der Untersuchungssituation (d.h., Verdunkelung, entspannter Untersuchungsgang für Kind und Untersucher) sollte das Spektrum der Hardware ausgeschöpft werden.

Klassischerweise werden zum Screening niederfrequente (um 3 MHz) Sektorschallköpfe verwendet. Ihr Vorteil liegt in der größeren Eindringtiefe und der Möglichkeit der Darstellung eines Organs in seiner Gesamtausdehnung. Details können dieser Untersuchungstechnik jedoch entgehen. Durch die konvexe Form des Schallkopfes trifft lediglich das zentrale Schallbündel im rechten Winkel auf das abzubildende Gewebe und ermöglicht eine unverzerrte 1:1 Darstellung.
Höherfrequente Linear-Schallköpfe (> 5 MHz) werden aufgrund ihrer geringeren Eindringtiefe üblicherweise für die Abbildung oberflächlicher Gewebe (Muskulatur, Sehnen) verwendet. Beim Kind eignen sie sich jedoch aufgrund des besseren Auflösungsverhaltens hervorragend auch für die Appendizitis-Diagnostik.
Es wurde die Hypothese aufgestellt, dass sich bei Kindern und schlanken Heranwachsenden die oberflächennahen Anteile der Milz mit höherfrequenten Schallköpfen besser als mit dem Standardverfahren darstellen lassen und so

[45] Stengel D. Plädoyer für mehr evidenzbasierte Chirurgie: Reizthema. *Dtsch Ärztebl* 2004;101:2398.

Kapselrisse und subkapsuläre Hämatome, die einen Risikofaktor für zweizeitige Rupturen darstellen, mit höherer Wahrscheinlichkeit ausgeschlossen werden können.

3.3.2. Ursachenforschung und Überlegungen zu technischen Modifikationen

3.3.2.1. Methoden

In eine Inzeptionskohortenstudie wurden zwischen Oktober 1998 und März 1999 kreislaufstabile Kinder ab 4 Jahre und schlanke Jugendliche unter 21 Jahren (max. Body Mass Index 22 kg/cm^2) eingeschlossen, die ein Hochrasanztrauma, eine Mehrfachverletzung oder ein stumpfes Bauchtrauma erlitten hatten. Es musste die Indikation zur kontrastmittelverstärkten Abdomen- oder Ganzkörper-CT gegeben sein, die den unabhängigen Referenzstandard dieser Studie darstellte.
Die Fallzahlplanung beruhte auf der Annahme einer 20% Verbesserung der Fläche unter der ROC-Kurve von 70% auf 90% durch den Einsatz des höherfrequenten Schallkopfes. Zur Erzielung einer Power von 80% mit einem zweiseitigen Testniveau von 5% sollten 34 Patienten eingeschlossen werden.

Nach der üblichen FAST-Untersuchung mit einem 3,5-MHz-Sektorschallkopf und Dokumentation des Befundes wurde eine nochmalige Sonografie der Milz mit einem 7,5-MHz-Linearschallkopf angeschlossen. Neben der Suche nach freier Flüssigkeit wurde das Parenchym mit beiden Methoden auf Unregelmäßigkeiten (echoreiche Kontusionen, echoarme Blutungen, Perfusionsausfälle) untersucht. Besonderes Augenmerk wurde auf Unterbrechungen der Kapsel gerichtet. Die Sonografien wurden ausschließlich durch zwei sehr erfahrene Untersucher durchgeführt.

Im Anschluss erfolgte in jedem Fall die Verifikation der Befunde durch CT. Der weitere klinische Verlauf wurde dokumentiert. Hauptzielkriterien waren jedoch die diagnostischen Testcharakteristika Sensitivität, Spezifität, Likelihood ratios und Flächen unter den ROC Kurven. Neben den Punktschätzern wurden 95% Konfidenzintervalle (KI) berechnet. Zur besseren Veranschaulichung der diskriminatorischen Wertigkeit wurden parametrische ROC-Kurven mit dem von Dorfman und Alf vorgeschlagenen Maximum-Likelihood-Schätzverfahren berechnet.[46]

---

[46] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *J Math Psychol* 1969;6:487-496.

3.3.2.2. Ergebnisse

Es wurden 37 Kinder (27 Jungen, 10 Mädchen) mit einem mittleren Alter von 17,8 ± Jahren und einem mittleren ISS von 38,2 ± 22,7 eingeschlossen (s. Tabelle 6). Die Prävalenz von Milzverletzungen betrug 54,1%.

	Keine Milzverletzung n=17	**Milzverletzung n=20**
Alter (Jahre)	18,7 (5,5 − 15,9)	17,0 (6,3 − 14,0)
Männliche Patienten (%)	70,6 (44,0 − 89,7)	75,0 (50,9 − 91,3)
Body Mass Index, BMI (kg/cm²)	19,9 (18,5 – 21,4)	19,8 (19,1 – 20,6)
Injury Severity Score, ISS	38,2 (28,5 − 47,9)	38,3 (26,1 − 50,4)
Beatmete Patienten (%)	47,1 (22,9 − 72,2)	70,0 (45,7 − 88,1)

Tabelle 6 Patientendemografie. Werte in Klammern entsprechen 95% Konfidenzintervallen.

Das Verletzungsmuster beinhaltete vier sofort operationsspflichtige Milzrupturen, neun Kapseleinrisse, vier subkapsuläre Hämatome und fünf Parenchymkontusionen. Aufgrund von zweizeitigen Rupturen bzw. einer Zunahme des Hämoperitoneums mit Kreislaufinstabilität mussten 13 Kinder operiert werden; ein milzerhaltendes Vorgehen gelang in 10 Fällen. Zwei Kinder verstarben innerhalb von 24 Stunden aufgrund von schweren Schädel-Hirn-Verletzungen.
Lediglich zwei Patienten fielen mit erheblichen intraabdominellen Flüssigkeitsmengen auf. Dreizehn Milzverletzte (65,0%, 95% KI 40,8 – 84,6%) zeigten nur Spuren eines Hämoperitoneums (s. Abbildung 14). Fünf Patienten (25,0%, 95% KI 8,7 – 49,1%) boten keinerlei Extravasat.

Im Vergleich zum Referenzstandard der CT lag die Sensitivität des 3,5-MHz-Schallkopfes bei 55,0% (95% KI 31,5 – 76,9%), diejenige des 7,5-MHz-Linearschallkopfes bei 90,0% (95% KI 68,3 – 98,8%). Beide Methoden lieferten einen falsch-positiven Befund (Spezifität 94,1%, 95% KI 71,3 – 99,9%).
Wurden Kapseleinrisse betrachtet (Prävalenz 35,1%), betrug die Sensitivität der beiden Schallköpfe 61,5% (95% KI 31,6 – 86,1%) und 92,3% (95% KI 63,9 – 99,8%). Die Spezifität lag mit beiden Methoden bei 95,8% (78,9 – 99,9%). Die Abbildung von Parenchymkontusionen (Prävalenz 13,5%) gelang mit einer Sensitivität von 40,0% (95% KI 5,3 – 85,3%) bzw. 80,0% (28,4 – 99,5%) und einer Spezifität von 96,9% (95% KI 83,8 – 99,9%).

Die Testcharakteristika beider Schallköpfe für die genannten Endpunkte sind in Abbildung 8 zusammengefasst.

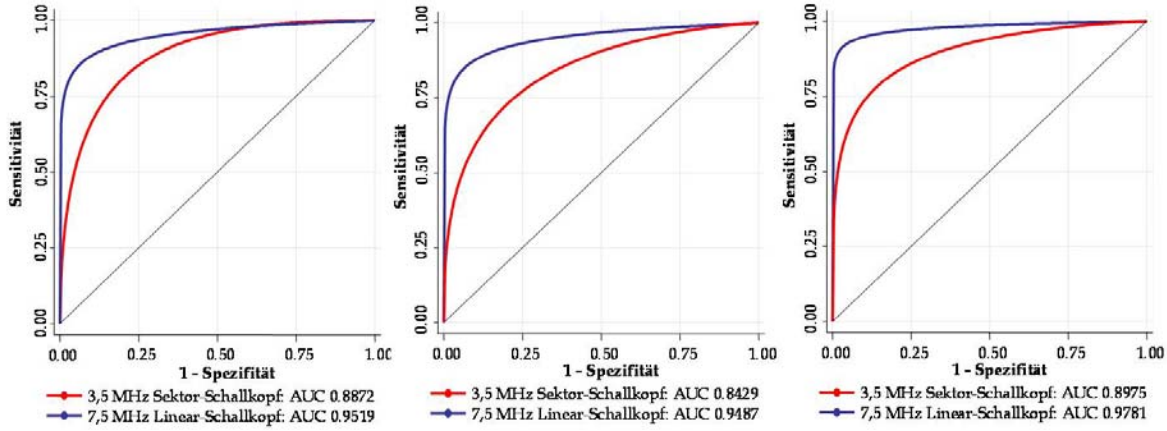

Abbildung 8 Testcharakteristika von 3,5-MHz-Sektor- und 7,5-MHz-Linearschallköpfen in der Diagnostik kindlicher Milzverletzungen im Vergleich zur kontrastmittelverstärkten Spiral-CT.
Links: alle Verletzungen. Mitte: Parenchymkontusionen. Rechts: Kapselrupturen.

3.3.2.3. Wissenschaftlicher Erkenntnisbeitrag und Einschränkungen der Interpretation

Die Ausschöpfung der apparativen Möglichkeiten und der differenzierte Einsatz verschiedener Schallköpfe kann die Sensitivität der Ultraschalluntersuchung in der Diagnostik kindlicher Milzverletzungen verbessern.
Die während der Planungsphase angenommenen Testcharakteristika ließen sich nicht reproduzieren. Die Sensitivität beider Schallköpfe fiel geringer als erwartet aus, die Flächen unter den ROC-Kurven waren größer als vermutet. Die Flächendifferenz betrug für unterschiedliche Endpunkte jedoch nur maximal 10%. Die mit dem 7,5-MHz-Schallkopf erzielte negative Likelihood Ratio von 0,16 führte nur zur moderaten Erniedrigung der Vortest-Wahrscheinlichkeit (von 54% auf 16%); das Ziel des *definitiven Ausschlusses einer Milzverletzung* konnte auch durch eine erweiterte Ultraschalluntersuchung nicht erreicht werden.
Aus klinischer Sicht wurde der diagnostische Zugewinn daher als nicht ausreichend beurteilt, um die Basisuntersuchung unnötig zu verzögern. Die geringe beobachtete Differenz rechtfertigte zudem keine weitere Studie mit Einschluss einer größeren Patientenstichprobe.

Die Untersuchung lieferte analog zu den Meta-Analyse-Daten Evidenz *gegen* den Einsatz der Ultraschalluntersuchung als diagnostisches Instrument zum Ausschluss kindlicher Milzverletzungen. Insbesondere bei hoher Vortest-Wahrscheinlichkeit (Prävalenz) sollte bei Kreislaufstabilität und entsprechender apparativer Ausstattung ein bildgebendes Referenzverfahren wie die CT angestrebt werden. Die differenzierte Sonografie kann im peripheren Krankenhaus ohne CT-Kapazität möglicherweise das diagnostische Instrumentarium bereichern; eine stationäre Überwachung ist jedoch auch bei negativem Ultraschallbefund angezeigt.

3.4. Forderung nach randomisierten Studien in der Diagnoseforschung: ein geeignetes Instrument zum Nachweis von Wirkungsgrad und Nutzwert

3.4.1. Kann ein hoher Nutzwert trotz mangelnder Wirksamkeit existieren?

Ohne Wirksamkeit kein Wirkungsgrad, ohne Wirkungsgrad auch keine Effizienz. Diese einfache Kosten-Nutzen-Bewertung wird im klinischen Alltag selten vorgenommen. Ein Test, der nicht messen kann, was er zu messen vorgibt, *sollte* keinen Einfluss auf Therapieentscheidungen nehmen. Der Arzt sieht sich täglich vor die schwierige Aufgabe gestellt

1) Informationen über die Effektivität eines Tests einzuholen,
2) diese Information formal zu bewerten (d.h., zu entscheiden, ob die in der Literatur berichteten Ergebnisse *valide* sind) und
3) zu entscheiden, ob die Ergebnisse des validen Tests seine Entscheidungen merklich und im Interesse seines Patienten oder der Gesellschaft beeinflussen werden.

Diese Aufgabe wird dem Kliniker derzeit nicht gerade leicht gemacht. Er steht im Spannungsfeld zwischen Traditionen, gesundem Menschenverstand, medizinischer Notwendigkeit, berechtigter und unberechtigter Anspruchshaltung von Patienten und deren Angehörigen, ökonomischen Grundsätzen und politischen Entscheidungen. Besondere Bedeutung erlangen diese häufig unvereinbaren Forderungen und Wünsche in der Screening-Problematik.

Mangelnde Effizienz eines Tests kann aus zwei Gründen resultieren

1) Die aus dem Testergebnis abgeleiteten medizinischen Interventionen haben keine Konsequenz auf das Outcome oder
2) Das Testergebnis selbst beeinflusst die ärztliche Entscheidung nicht oder nur unwesentlich.

Offensichtlich lassen sich diese unterschiedlichen Bedingungen nur schwer voneinander trennen. Bei einer konventionellen ökonomischen Betrachtung (z.B. einer cost-utility-Analyse) steht bereits das Zielkriterium eines Gewinns für den Patient (gemessen z.B. in Quality-adjusted life years [QALY]) oder die Gesellschaft im Vordergrund.
Eine Unterscheidung ist jedoch notwendig, da ein Test mit geringem Wirkungsgrad bereits in einer sehr viel früheren Phase seiner klinischen Erprobung verworfen werden sollte.
Die nach Veröffentlichung der systematischen Übersichtsarbeit geführte weltweite Korrespondenz hat gezeigt, dass die postulierte These „ohne Effektivität keine Effizienz“ nicht ohne weiteres in der wissenschaftlichen Gemeinschaft akzeptiert wird.
Die Reaktionen auf die Ergebnisse der Meta-Analyse waren erwartungsgemäß zwiespältig- Kritiker Ultraschall-basierter Schockraum-Algorithmen sahen sich in ihrer Auffassung bestärkt, Befürworter zweifelten die Interpretation der Daten an.

Als wichtigstes Argument wurde angeführt, dass die *Bestätigung* einer klinisch vermuteten Abdominalverletzung die weiteren therapeutischen Entscheidungen relevant beeinflusst, insbesondere jedoch die Zeit bis zu einer Laparotomie verkürzt. Auch wurde angemerkt, dass die hohe Spezifität der Sonografie das Risiko für eine negative (also unnötige, potenziell jedoch komplikationsbehaftete) Laparotomie erniedrigen würde.
Amerikanische Kollegen wiesen auf die Kostenreduktion hin, die aus der Vermeidung von CT und stationärer Überwachung resultierten. Der Einwand, dass die verfügbaren Daten (und die offensichtlich geringe Sensitivität der Sonografie) gerade den Verzicht auf die CT oder stationäre Überwachung *nicht* erlauben würden, wurde mit dem Verweis auf vergleichende Interventionsstudien abgelehnt.

Für den Nachweis der Effektivität (also diagnostischen Genauigkeit) einer Methode ist die unabhängige Prüfung des Testergebnisses durch einen Referenzstandard am gleichen Individuum zwingend erforderlich. Die randomisierte Studie (die zwei sich in ihren biologischen Eigenschaften ähnliche Populationen vergleicht) kann keine Prüfung der Effektivität leisten, aber Wirkungsgrad und Nutzwert demaskieren.

Hierzu werden Patienten zufällig Gruppen zugeordnet, die mit dem experimentellen oder einem etablierten (bzw. dem experimentellen und etablierten oder lediglich dem etablierten) Test untersucht werden. Alternativ können auch Ärzte randomisiert werden, in Kenntnis oder Unkenntnis des Testergebnisses zu handeln. Eine derartige Studie macht Sinn, wenn der interessierende Test seine Effektivität unter Beweis gestellt hat.
Eine randomisierte Studie mit und ohne Ultraschalluntersuchung erscheint daher vor dem Hintergrund der gezeigten Ergebnisse widersinnig- dennoch wurden derartige Experimente durchgeführt.

Um den kritischen Stimmen gerecht zu werden, lag es nahe, auch eine systematische Übersicht und Meta-Analyse randomisierter Studien durchzuführen, die als Cochrane-Review realisiert werden konnte. Die Cochrane-Gesellschaft akzeptiert als Datenquellen derzeit ausschließlich randomisierte (randomized controlled trial, RCT) und sog. quasi-randomisierte Studien (qRCT, mit einer Zuteilung der Teilnehmer z.B. nach Aufnahmenummer oder Geburtsdatum).

3.4.2. Identifikation und Summierung randomisierter und quasi-randomisierter Studien zum Ultraschall-Problem

3.4.2.1. Methoden

Für diese systematische Übersicht wurden in Medline, Embase, CENTRAL, Cinahl, zahlreichen Verlagsdatenbanken sowie dem Internet RCT und qRCT identifiziert.

Zusätzliche wurden führende Zeitschriften, die Beiträge über die Schockraum-Sonografie enthielten, manuell auf potenziell relevante Arbeiten hin durchgesehen. Die Suche umfasste auch Abstract-Bände wissenschaftlicher Kongresse. Die Autoren wurden kontaktiert und um unpublizierte Informationen gebeten; der vertrauliche Umgang mit den Daten wurde zugesichert.

Für die Synthese relativer Risiken, Risiko- und Mittelwertdifferenzen wurden nach Heterogenitäts-Prüfung fixed- oder random-effects-Modelle (RevMan-Software 4.2 der Cochrane Collaboration) verwendet.

3.4.2.2. Ergebnisse

Von 377 Zitaten erwiesen sich lediglich fünf Arbeiten als RCT bzw. qRCT. Eine dieser Arbeiten wurde bisher lediglich als Kongress-Abstract publiziert; eine Anfrage

bei den Autoren blieb unbeantwortet. Ein RCT wurde konzipiert, um die *Überlegenheit der frühen CT-Untersuchung* gegenüber multi-interventionellen Ansätzen (Sonografie, DPL) zu überprüfen. Trotz zweiseitiger Prüfung sind die Ergebnisse dieser Untersuchung aufgrund der Vertauschung von experimenteller und Kontroll-Intervention in diesem Kontext daher mit Vorsicht zu interpretieren.
Der Versuch der Kontaktaufnahme mit einem dritten Autor scheiterte aufgrund ungewöhnlicher Gründe- er befindet sich wegen Betruges in fünfjähriger Haft. Ihm wurde die Approbation aberkannt und eine Geldstrafe von 1,5 Millionen Dollar auferlegt.

Die Ergebnisse der quantitativen Analyse dichotomer Endpunkte sind in Abbildung 10 dargestellt.
Ein statistisch signifikanter Effekt konnte lediglich für die Häufigkeit von CT-Untersuchungen nachgewiesen werden- bei jedem zweiten Patienten, der mit Hilfe eines ultraschallbasierten im Vergleich zu einem Kontroll-Algorithmus ohne Ultraschalluntersuchung diagnostiziert wurde, wurde auf eine CT-Untersuchung verzichtet (RD 46%, 95% KI 13 – 100%). Es bleibt jedoch unklar, ob dieser Surrogat-Parameter als Vorteil oder Nachteil interpretiert werden muss. Trotz Trends in den Punktschätzern ließen sich die postulierten Vorteile der Schockraum-Sonografie (Reduktion der negativen Laparotomieraten) nicht nachweisen. Argumentiert man mit Trends, muss fairerweise auch auf ein bedenkliches Ergebnis hingewiesen werden- das erhöhte Sterberisiko im Ultraschall-Arm (Risikodifferenz 11%, 95% KI -6 – 28%). Die wenigen verfügbaren Studien lassen jedoch kaum belastbare Rückschlüsse hinsichtlich aller interessierenden Endpunkte zu.

Ultraschall-basierte Algorithmen führten zu einer durchschnittlichen Verkürzung der Zeit bis zum Abschluss der Schockraum-Diagnostik von 98 Minuten (95% KI -122 – 74 Minuten).
In einer Studie wurde in der Ultraschall-Gruppe die Liegedauer auf der Intensivstation im Mittel um 0,6 Tage (95% Konfidenzintervall -0,9 – 2,1 Tage) reduziert; ein Einfluss auf die Hospitalisierungszeit ließ sich jedoch nicht nachweisen (WMD 0,0 Tage, 95% KI -2,2 – 2,2 Tage).
Hinsichtlich der direkten Kosten ergab sich kein einheitliches Bild. In der Untersuchung von Boulanger wurden Einsparungen von 384 US$ (95% KI 358 – 411 US$) pro Patient erzielt. Navarrete-Navarro hingegen fand in dieser Gruppe zusätzliche Aufwendungen von 148 US$ (95% KI 38 – 257 US$)- zu unterstreichen sind die gegenläufigen Nullhypothesen der genannten Untersuchungen.

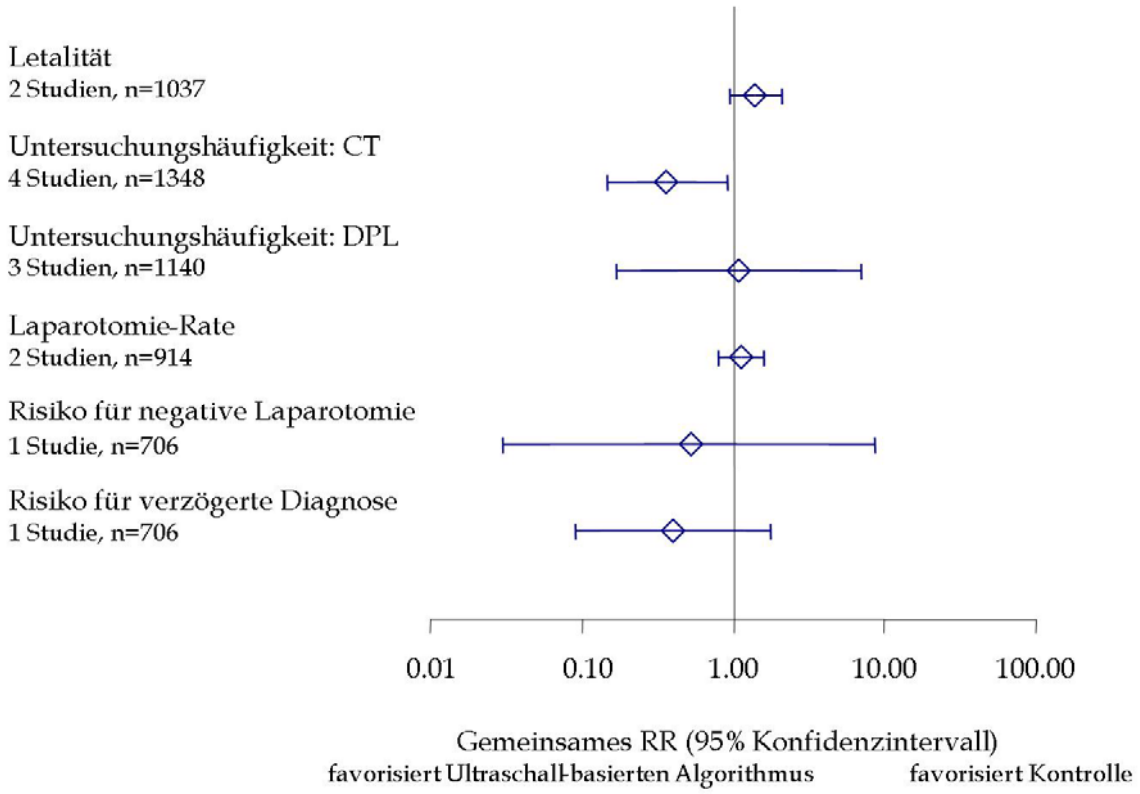

Abbildung 10 Ergebnisse des Cochrane-Reviews.

3.4.2.3. Wissenschaftlicher Erkenntnisbeitrag und Einschränkungen der Interpretation

Es konnte gezeigt werden, dass

1) die angenommene Wirksamkeit einer diagnostischen Methode aus einer selektiven Betrachtung der Referenzliteratur resultieren kann. Erst ihre systematische Aufarbeitung ist in der Lage, eine unzureichende Genauigkeit zu demaskieren.
2) die methodische Qualität einen erheblichen Einfluss auf die berichtete Genauigkeit nimmt. Hierbei tendieren Studien, die z.B. dem Prinzip einer unabhängigen Bestätigung des Testergebnisses durch einen Referenztest nicht gerecht werden, zu einer Überschätzung der Sensitivität. Instrumente wie QUADAS und STARD stellen gute Orientierungshilfen dar, um die Bewertung der Validität der erhobenen Daten zu systematisieren.
3) eine als Reaktion auf eine „negative“ Meta-Analyse Alltag durchgeführte klinische Untersuchung z.B. durch Modifikation technischer Details die Reputation eines Tests unter bestimmten Fragestellungen erhöhen kann. Meta-

Analysen können nicht alle klinisch relevante Szenarien abbilden. Eine in einer Meta-Analyse vernachlässigte oder unzureichend gewürdigte Situation, in der der untersuchte Test möglicherweise hilfreich ist, erfordert daher die erneute Durchführung einer Studie. Diese Bewertung muss jedoch durch den Kliniker erfolgen.

4) ein zweifelhaft effektives Verfahren in RCT erwartungsgemäß auch keinen günstigen Wirkungsgrad bzw. Nutzwert demonstrieren kann. Die ersten Meta-Analyse-Daten lagen im Mai 2001 vor und verletzten damit bereits das Prinzip der therapeutischen Unsicherheit (Equipoise).[47] Es existiert keine Rationale für weitere randomisierte Studien zu diesem Thema.

Der Nachweis der Effektivität (also der diagnostischen Genauigkeit) eines Tests ist seine Eintrittskarte in den Versorgungsalltag. Leider erweisen sich viele Diagnoseverfahren nach Bewältigung dieser Hürde als Danaergeschenk[48]- sind sie erst in der klinischen Praxis etabliert, werden sie nur schwer wieder verlassen. Einer hohen Testgenauigkeit wird derzeit noch ein höherer Stellenwert als dem Wirkungsgrad und der Effizienz beigemessen.

---

[47] Stengel D, Bauwens K, Ekkernkamp A. Unfallchirurgische Interventionsstudien: randomisiert oder nicht-randomisiert? *Unfallchirurg* 2003;106:194-199.

[48] Equo ne credite, Teucri! Quidquid id est, timeo Danaos et dona ferentes. (*Traut nicht dem Pferd, Trojer! Was auch immer es ist, ich fürchte die Danaer, auch wenn sie Geschenke bringen*.).Vergil, *Aenei*s