ORIGINAL ARTICLE

Accuracy of ultrasound to identify chronic liver disease

Richard Allan, Kerry Thoirs, Maureen Phillips

Richard Allan, Division of Medical Imaging, Flinders Medical Centre, Flinders Drive, Bedford Pk, 5042, South Australia, Australia

Richard Allan, Maureen Phillips, School of Health Sciences, University of South Australia, City East Campus, North Terrace, Adelaide, 5000, South Australia, Australia

Kerry Thoirs, International Centre of Allied Health Evidence, University of South Australia, City East Campus, North Terrace, Adelaide, 5000, South Australia, Australia

Correspondence to: Dr. Kerry Thoirs, International Centre of Allied Health Evidence, University of South Australia, City East Campus, North Terrace, Adelaide, 5000, South Australia, Australia. kerry.thoirs@unisa.edu.au

Telephone: +61-8-83022903 Fax: +61-8-83021818

Abstract

AIM: To identify and assess studies reporting the diagnostic performance of ultrasound imaging for identifying chronic liver disease (CLD) in a high risk population.

METHODS: A search was performed to identify studies investigating the diagnostic accuracy of ultrasound imaging for CLD. Two authors independently used the quality assessment of diagnostic accuracy studies (QUADAS) checklist to assess the methodological quality of the selected studies. Inter-observer reliability of the QUADAS tool was assessed by measuring the degree of agreement (percent agreement, κ statistic) between the reviewers for each assessment prior to a consensus meeting. The characteristics of each study population, sensitivity and specificity results for the index tests, and results of any testing for observer agreement were extracted from the reports. Receiver Operator Characteristic plots were generated using Microsoft Excel 2003 software and used to graphically display the diagnostic performance data and to explore the relationships between the reported ultrasound techniques and study characteristics, and methodology quality.

RESULTS: Twenty-one studies published between 1991 and 2009 were retained for data extraction, analysis and assessment for methodological quality. Assessment of methodology quality was performed on the 21 selected studies by two independent reviewers (RA & KT) using the QUADAS assessment tool. Across all studies the mean number of responses within the QUADAS assessment tool was 10 (range 7-13) for "Yes", 1 (range 0-3) for "No" and 3 (range 0-6) for "unclear". Inter-rater agreement for assessment of methodology quality was significantly greater than chance when assessing for representative spectrum, clear selection criteria, appropriate delay between reference and index tests, adequate descriptions of the index and reference tests, reference and index test blinding, and if relevant clinical information was provided. Seven studies reported moderate to high observer agreement for ultrasound techniques. Studies which clearly reported blinding performed better than the other studies for diagnostic accuracy, and lower diagnostic accuracy was evident for populations with lower prevalence of disease. Assessment of the liver surface using ultrasound consistently had moderate diagnostic accuracy across studies which demonstrated good research methodology. Other techniques demonstrated variable or poor to fair diagnostic accuracy.

CONCLUSION: Ultrasound of the liver surface is a useful diagnostic tool in patients at risk of CLD when assessing whether they should undergo a liver biopsy.

INTRODUCTION

Chronic liver disease (CLD) is a significant cause of morbidity and mortality in developed nations. It is commonly caused by viral hepatitis and alcohol abuse with significant contributions from metabolic disorders[1]. Accurate diagnostic testing for CLD to identify asymptomatic patients in a high risk population has become more important due to recent advances in management and treatment options that provide better patient outcomes if the diagnosis of fibrosis or cirrhosis can be made before cirrhosis becomes clinically apparent[2]. In some cases, liver fibrosis has been demonstrated to be reversible[3], a phenomenon that was previously not considered possible.

The standard method for determining, staging and grading CLD is liver biopsy[4]. The invasiveness of this method, and its associated morbidity and mortality has led to the emergence of less invasive methods which include medical imaging techniques (computed tomography, magnetic resonance imaging and ultrasound), serum markers (both direct and indirect markers of fibrosis) and transient elastography[2]. All of these techniques have the potential to reduce the number of biopsies performed in a high risk population.

Ultrasound can identify the manifestations of CLD such as liver fibrosis and cirrhosis which are characterized by the presence of vascularized fibrotic septa and regenerating nodules[1,5-7]. Ultrasound is an attractive diagnostic tool because it is readily available, inexpensive, well tolerated and is already extensively used in the diagnostic work-up of patients with CLD. The diagnostic accuracy of ultrasound needs to be established to inform clinicians of its role in patients at high risk of CLD.

The aim of the following systematic review was to identify and assess studies reporting the diagnostic performance of ultrasound imaging for identifying CLD in a high risk population.

MATERIALS AND METHODS

Search strategy

A search of electronic databases in November 2009 was performed by one author (RA) to identify studies reported in English, investigating the diagnostic accuracy of ultrasound imaging for CLD. MEDLINE, EMBASE, CINAHL and Science Citation Index databases were searched using the terms "chronic liver disease", "cirrhosis", "fibrosis", "liver biopsy". The truncated terms "sonograph*" and "ultraso*" were also used in the search for alternate terms used for ultrasound such as sonography, sonographic, ultrasonic, ultrasound and ultrasonography. A Boolean search strategy was employed for the above terms in the following form: (sonograph* OR ultraso*) AND (chronic liver disease OR cirrhosis OR fibrosis) AND liver biopsy. No search filters were used. "Pearling" of the reference lists of all selected studies was also performed.

Eligibility and study selection

One author (RA) determined the eligibility of studies for inclusion in this review. Inclusion and exclusion criteria were created to identify studies that were likely to conform to the highest level of evidence for studies of diagnostic tests using the National Health and Medical Research Council of the Australian Government Level II criteria[8].

The inclusion and exclusion criteria for the systematic review are described in Table 1. Initially, abstracts of all identified studies were assessed to determine if the study met the inclusion and exclusion criteria. Studies were retained if they clearly met the inclusion criteria, did not meet the exclusion criteria, or if it was unclear from the abstract if the study met the exclusion and inclusion criteria. The full text reports of all retained studies were then re-assessed for inclusion. All studies clearly meeting any of the exclusion criteria were excluded, and all studies meeting all the inclusion criteria were retained for assessment of methodological quality, data extraction and analysis.

Assessment of methodological quality

Two authors (RA, KT) independently used the quality assessment of diagnostic accuracy studies (QUADAS)[9] checklist to assess the methodological quality of the selected studies. The QUADAS checklist (Table 2) contains 14 assessment items, each assessing an aspect of the study that impacts on methodological quality. Each author assessed the selected studies by rating each assessment item for each study as "yes", "no" or "unclear". The studies were not given an overall score, nor were they stratified into high or low quality groups. Inter-observer reliability of the QUADAS tool was assessed by measuring the degree of agreement (percent agreement, κ statistic) between the reviewers for each assessment prior to a consensus meeting. A consensus meeting was held to resolve any discrepant scores between the two assessors. A third independent assessor (MP) reviewed the discrepant scores and acted as a final adjudicator if a consensus could not be reached.

Data extraction

The characteristics of each study population were extracted from the reports and included country of origin, sample size, gender, aetiology, age (mean, range and SD),

Table 1 Inclusion and exclusion criteria for studies

Inclusion criteria	Exclusion criteria
Evaluated diagnostic accuracy	Did not evaluate diagnostic accuracy
Quantitative results of diagnostic performance presented in a format that enabled a 2 × 2 contingency table to be extracted OR results presented as sensitivity, specificity and prevalence	2 × 2 contingency table could not be extracted from results of diagnostic performance OR sensitivity, specificity and prevalence results not presented
Index test of study was an ultrasound imaging technique	Index test included was not an ultrasound imaging technique OR included a non-ultrasound imaging technique as part of the index test
Studies were conducted prospectively	Studies were not conducted prospectively
The reference test for all subjects in the study was liver biopsy	The reference test for the study was not liver biopsy OR liver biopsy was not used for all subjects
The sample population described were adults at risk of chronic liver disease	The sample population described included children OR sample population included adults not at risk of chronic liver disease
	The study was published as a case study, review or editorial

Table 2 Quality assessment of diagnostic accuracy studies assessment items

Item	Question	Guidelines for assessment	Aspect of study assessed
1	Was the spectrum of patients representative of the patients who will receive the test in practice?	Patients who receive the test in clinical practice will be suspected of having chronic liver disease but not yet have decompensated cirrhosis. Sample populations should fit this general characteristic. Samples may be a mixed population or may be restricted to one disease type if this is a common and clinically important disease, in this case alcohol abusers or viral hepatitis. Score "yes" if clearly stated and meet the above definitions, "no" if the spectrum is clearly outside this definition and "unclear" if there is insufficient information	Generalisability
2	Were selection criteria clearly described?	Clear definitions of the inclusion and exclusion criteria should be included. "Yes" if clearly stated, "no" if not stated and 'unclear' if only partially stated	Quality of reporting
3	Is the reference standard likely to correctly classify the target condition?	Liver biopsy must be used as the reference standard. "Yes" if biopsy used, "no" if not and "unclear' if not stated	Presence of bias
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	The time period must be no more than one month for all cases to avoid discrepancies between the index and reference test due to disease progression. The order in which the tests are done is not relevant. Score "yes" if one month or less, "no" if more than one month and "unclear" if not clearly stated	Presence of bias
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard?	All patients should receive a biopsy unless some form of randomisation was used. Score "no" if some patients were excluded. Score "unclear" if this information is not reported by the study	Presence of bias
6	Did patients receive the same reference standard regardless of the index test result?	If it is clear all patients received a liver biopsy, score "yes". If some received laparoscopy (or other test), score "no". If it is not stated, score "unclear"	Presence of bias
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	Score 'yes' if the index test did not form part of the reference test, "no" if it did and "unclear" if not stated or there is doubt	Presence of bias
8	Was the execution of the index test described in sufficient detail to permit replication of the test?	Studies should describe equipment and techniques in sufficient detail to enable replication. Ultrasound criteria for identifying fibrosis or cirrhosis must be clearly stated and be able to be replicated (e.g. clear and easily reproducible system for assessing grey scale appearances or Doppler measurements or indices). Score "yes" if the above is true, "no" if these details are not stated or if the technique described is not able to be replicated and "unclear" if an incomplete description is given	Quality of reporting
9	Was the execution of the reference standard described in sufficient detail to permit its replication?	A clear description of the biopsy technique sufficient to enable replication. Ideally this should include information about the needle technique used and the minimum size of the sample. A recognised staging system for fibrosis or a description with sufficient detail to enable replication must be provided. Score "yes" if the above are true, "no" if no description of technique is given OR no staging system used and "unclear" if a partial description is given from which conclusions cannot be reached	Quality of reporting
10	Were the index test results interpreted without knowledge of the results of the index test?	Score "yes" if the ultrasound was performed and reported without knowledge of the biopsy. Score "no" if this is not the case and "unclear" if it is not stated	Presence of bias
11	Were the reference standard results interpreted without knowledge of the results of the reference test?	Score "yes" if the biopsy was performed and reported without knowledge of the ultrasound. Score "no" if this is not the case and "unclear" if it is not stated	Presence of bias

12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	Score "yes' if pre-test clinical data was available for the ultrasound and biopsy. Score "no" if it was not available. Score "unclear" if it is not stated	Presence of bias
13	Were uninterpretable/intermediate results reported?	Score "yes" if all test results, including uninterpretable or indeterminate results, are accounted for. Score "no" if some data is missing and not explained or has been excluded from analysis. Score "unclear" if it is not clear whether all results have been included	Quality of reporting
14	Were withdrawals from the study explained?	A flow chart or matching numbers in a 2 × 2 table can help assess this item. If it is clear what happened to all participants, score "yes". If some patients are not accounted for, score "no". Score "unclear" if interpretation is difficult	Quality of reporting

Table 3 Inter-rater reliability for quality assessment of diagnostic accuracy studies items

QUADAS item	Agreement (%)	κ
Representative spectrum?	90	0.462[1]
Selection criteria clear?	81	0.632[1]
Appropriate reference standard?	100	1.000[1]
Appropriate delay between tests?	100	1.000[1]
Partial verification avoided?	95	-[2]
Differential verification avoided?	95	-[2]
Incorporation avoided?	100	1.000[1]
Adequate index test description?	86	0.468[1]
Adequate reference test description?	76	-[2]
Index test blinded?	86	0.704[1]
Reference test blinded?	95	0.901[1]
Relevant clinical information available?	86	0.712[1]
Uninterpretable results reported?	29	0.022
Withdrawals explained?	33	0.033

[1]Agreement significantly greater than chance ($P < 0.05$); [2]A κ statistic could not be calculated because one reviewer responded "yes" for all studies on this item. QUADAS: Quality assessment of diagnostic accuracy studies.

exclusion and inclusion criteria, severity of disease, prevalence, staging system of liver biopsy, and the ultrasound technique(s) used. Sensitivity and specificity results for the index tests were extracted from the reports or from constructed contingency tables. The results of any testing for observer agreement were also extracted.

Statistical analysis
Receiver Operator Characteristic (ROC) plots were generated using Microsoft Excel 2003 software and used to graphically display the diagnostic performance data and to explore the relationships between the reported ultrasound techniques and study characteristics[10]. To demonstrate any patterns and relationships between methodology quality and diagnostic quality, plots were also produced for items on the QUADAS checklist.

RESULTS

Search results
No previous systematic reviews addressing the diagnostic accuracy of ultrasound in liver fibrosis or cirrhosis were identified. A total of 1355 separate studies were revealed from the following databases: MEDLINE ($n = 464$), EMBASE ($n = 1155$), CINAHL ($n = 18$) and Science Citation Index searches ($n = 639$). Attrition of studies after

an initial assessment of the abstracts against the inclusion and exclusion criteria resulted in a residual of 38 studies [MEDLINE ($n = 33$), EMBASE ($n = 3$), Science Citation Index ($n = 2$)]. An additional 8 studies were revealed after pearling of the residual 38 studies ($n = 46$). After assessment of the full text reports of these 46 studies against the selection criteria, there was further attrition of 25 studies resulting in a total of 21 studies retained for data extraction, analysis and assessment for methodological quality.

Methodology quality assessment results
Assessment of methodology quality was performed on the 21 selected studies by two independent reviewers (RA & KT) using the QUADAS assessment tool. Inter-rater agreement for each item, across all studies, was assessed by calculating the percentage agreement and kappa value (κ) (Table 3). For items where there was disagreement between the reviewers, consensus was achieved without the need for an independent adjudicator.

Across all studies the mean number of responses within the QUADAS assessment tool was 10 (range 7-13) for "Yes", 1 (range 0-3) for "No" and 3 (range 0-6) for "unclear".

Characteristics of study populations
The studies included in this review were published between 1991 and 2009. The characteristics of the study populations are reported in Table 4.

The method for staging the histology obtained at liver biopsy was either not reported or unclear in 5 studies, all of which were published prior to the year 2000. Across the other 16 studies a total of seven staging systems were used. METAVIR[11] ($n = 7$), Ishak[12] ($n = 3$), Desmet[13] ($n = 2$) and four other systems which were each used once[14-17].

Measurements of observer agreement
Seven studies reported observer agreement assessment of the ultrasound technique[18-24]. When reported, results for observer agreement were acceptable, with κ values ranging from 0.51-0.93, coefficient of variation values ranging from 2%-8%, and correlation coefficients ranging from 0.82-0.9.

Ultrasound techniques
Diagnostic accuracy was determined for a range of ultra-

Table 4 Characteristics of included studies

Author	Country	Sample	Males (%)	Mean age in years (range)	Prevalence of disease (%)	Aetiology (largest disease type)	Inclusion criteria	Exclusion criteria	Severity of disease
Joseph et al[17]	UK	50	NR	NR (NR)	62	Mixed (alcohol)	Abnormal LFT, clinical suspicion	NR	NR
Cioni et al[32]	Italy	117	77 (66)	47 (NR)	50	NR	Raised ALT	Decompensation, refused biopsy	Mild
Ladenheim et al[26]	USA	50	NR	NR (NR)	16	NR	NR	NR	NR
Ferral et al[35]	Mexico	70	28 (40)	49 (18-84)	46	Unclear	Abnormal LFT, non-specific clinically	Did not have biopsy (reasons not specified)	NR
Hultcrantz et al[28]	Sweden	83	47 (57)	41 (NR)	17	Mixed ("fatty" 54%)	Asymptomatic, raised AST/ALT	Signs of liver disease	Mild
Colli et al[29]	Italy	52	30 (58)	52 (22-65)	31	Viral	HCV, Child-Pugh class "A"	Decompensation, PHT	Mild
Gaiani et al[20]	Italy	212	128 (60)	49 (15-71)	22	Mixed (HCV 57%)	Raised AST, no prev. cirrhosis	Decompensation, PHT, previous history cirrhosis	Mild
Xu et al[22]	China	66	42 (64)	39 (NR)	36	Viral	HBV	NR	NR
Mathiesen et al[27]	Sweden	165	110 (67)	48 (22-77)	9	Mixed ("fatty" 40%)	Asymptomatic, raised AST/ALT	Decompensation	Mild
Colli et al[18]	Italy	300	234 (78)	49 (17-78)	36	Mixed (HCV 41%)	Asymptomatic, raised AST/ALT	Heart failure, atrial fibrillation	Mild
Nishiura et al[25]	Japan	103	60 (58)	51 (38-75)	21	Mixed (viral 88%)	Raised AST, no prev. cirrhosis	Decompensation, previous history cirrhosis	Mild
Colli et al[19]	Italy	176	96 (55)	54 (NR)	38	Viral	HCV, raised AST, Child-Pugh "A"	Decompensation, biopsy contra-indicated	Mild
Vigano et al[33]	Italy	108	55 (51)	53 (NR)	34	Viral	HCV	NR	NR
D'Onofrio et al[31]	Italy	105	73 (70)	47 (NR)	27	Viral	Asymptomatic viral hepatitis, raised AST/ALT	NR	Mild
Schneider et al[30]	Germany	119	66 (55)	45 (20-78)	14	Viral	HCV	NR	NR
Shen et al[16]	China	324	272 (84)	36 (18-60)	9	Viral	HCV,HBV, raised ALT	Decompensation, HIV, other causes of CLD	Mild
Liu et al[21]	Taiwan	503	271 (54)	52 (NR)	33	Viral	HCV	HBV, HIV, NASH, alcohol abuse, refused biopsy or contra-indicated	NR
Iliopoulos et al[23]	Greece	72	45 (63)	57 (NR)	39	Viral	Unclear	Unclear	NR
Paggi et al[24]	Italy	430	237 (55)	53 (25-71)	37	Viral	HCV	HBV, HIV, decompensation	Mild
Wang et al[39]	Taiwan	320	199 (62)	51 (NR)	33	Viral	HBV, HCV	HCC	NR
Gaia et al[34]	Italy	61	41 (67)	NR	36	Viral (62%)/NASH (38%)	NR	NR	NR

LFT: Liver function test; NR: Not reported; AST: Aspartate aminotransferase; ALT: Alanine aminotransferase; PHT: Portal hypertension; CLD: Chronic liver disease; HIV: Human immunodeficiency virus; NASH: Non-alcoholic steato-hepatitis; HCV: Hepatitis C virus; HBV: Hepatitis B virus; HCC: Hepatocellular carcinoma.

sound techniques across all studies. There were 48 reports of diagnostic accuracy for specific ultrasound techniques within the 21 included studies. Thirty different ultrasound techniques were reported of which 23 were reported once. Seven techniques were reported multiple times. The ultrasound techniques could be broadly described according to four main categories: (1) low frequency grey scale imaging, where an assessment of the liver parenchyma, liver shape and size, spleen size and hepatic vessel appearance or calibre was made from an ultrasound examination using a low frequency ($\leqslant$ 5 MHz) convex or sector transducer (n = 14 reports); (2) high frequency grey scale imaging, where the liver surface was assessed using a high frequency linear (> 5 MHz) array transducer (n = 8 reports); (3) Doppler techniques, where a Pulsed Wave (PW) Doppler study of the portal, hepatic and splenic veins and/or the hepatic artery was performed to determine measurements of maximum or mean velocities, ratios

and/or indices of resistance and/or pulsatility, and/or subjective assessments of haemodynamic waveforms (n = 19 reports); and (4) Scoring system using a combination of techniques, where more than one technique and/or parameter described in categories 1-3 provided a quantitative or qualitative assessment (n = 7 reports).

The diagnostic accuracy of the ultrasound techniques by group are demonstrated in Table 5.

Statistical analysis

A ROC plot (Figure 1A) was generated for all 48 reports of diagnostic accuracy according to the predetermined broad group categories. One scoring system achieved perfect results[25], while one report of high frequency liver surface technique[26] indicated a performance no better than chance.

A ROC plot (Figure 1B) was generated for ultrasound techniques that were reported more than once.

Table 5 Diagnostic accuracy of all ultrasound techniques

Study	Specific technique	Sensitivity	Specificity
Low frequency grey scale techniques			
Schneider *et al*[30]	Spleen width	86.3	35.3
Schneider *et al*[30]	Spleen length	77.5	53.0
Joseph *et al*[17]	Liver parenchyma heterogeneity	77.0	89.0
Shen *et al*[16]	PV diameter	76.7	45.0
Iliopoulos *et al*[23]	Spleen volume	75.0	70.0
Shen *et al*[16]	Spleen length	60.0	75.0
Shen *et al*[16]	Splenic vein diameter	60.0	78.0
Hultcrantz *et al*[28]	Liver parenchyma echogenicity	43.0	42.0
Iliopoulos *et al*[23]	Liver parenchyma heterogeneity	43.0	77.0
Colli *et al*[18]	Caudate/Right lobe ratio	41.0	91.0
Mathiesen *et al*[27]	Liver parenchyma echogenicity	40.0	38.6
D'Onofrio *et al*[31]	Collateral vessels	39.0	84.0
D'Onofrio *et al*[31]	Caudate/Right lobe ratio	32.0	99.0
D'Onofrio *et al*[31]	Liver parenchyma heterogeneity	29.0	99.0
High frequency grey scale techniques			
Ferral *et al*[35]	Surface	87.5	81.6
Colli *et al*[19]	Surface	60.0	92.0
Colli *et al*[18]	Surface	54.0	95.0
D'Onofrio *et al*[31]	Surface	54.0	78.0
Vigano *et al*[11]	Surface	51.0	90.0
Ladenheim *et al*[12]	Surface	12.5	88.0
Gaia *et al*[34]	Surface	63.0	86.0
Paggi *et al*[24]	Surface	73.0	90.0
Doppler techniques			
Liu *et al*[21]	SA PI = 0.85	94.0	39
Liu *et al*[21]	SA PI = 1.20	88.0	82
Iliopoulos *et al*[23]	PV congestion index (PV cross-sectional area/PV Vtam)	86.0	66
Iliopoulos *et al*[23]	PV Diameter/PV Vmax	86.0	59.0
Iliopoulos *et al*[23]	PV Diameter/Vtam	86.0	68.0
Iliopoulos *et al*[23]	HA Vtam/PV Vtam	86.0	61.0
Iliopoulos *et al*[23]	PV Vmax	77.0	71.0
Schneider *et al*[30]	PV undulations	76.5	100.0
Colli *et al*[29]	HV pulsatility	75.0	78.0
Iliopoulos *et al*[23]	PV Vtam	75.0	71.0
Schneider *et al*[30]	PV Vmax	74.5	53.0
Iliopoulos *et al*[23]	HA RI	71.0	55.0
Cioni *et al*[32]	PV Vmax	66.0	98.0
Liu *et al*[21]	SA PI = 1.10	61.0	98.0
Iliopoulos *et al*[23]	PV blood flow (BF) (mL/min)	59.0	75.0
Colli *et al*[18]	HV pulsatility	57.0	76.0
Liu *et al*[21]	SA PI = 1.40	45.0	99.0
Iliopoulos *et al*[23]	Doppler perfusion index HA BF/(HA BF + PV BF)	43.0	91.0
Schneider *et al*[30]	HV pulsatility	31.4	47.1
Scoring systems			
Nishiura *et al*[25]	Sequential score (high and low frequency techniques)	100.0	100.0
Xu *et al*[22]	4 parameter score (low frequency techniques)	87.8	97.6
Gaiani *et al*[20]	Score of low frequency and PV Vtam	82.2	79.9
Gaiani *et al*[20]	Score of 5-7 techniques (low frequency and PV Vtam)	78.7	80.6
D'Onofrio *et al*[31]	Any of 4 techniques (low frequency and liver surface)	68.0	68.0
D'Onofrio *et al*[31]	All of 4 techniques (low frequency and liver surface)	25.0	100.0
Wang *et al*[39]	Score of 4 parameters (low frequency techniques)	74.0	86.0

PV: Portal vein; SA: Splenic artery; PI: Pulsatility index; Vtam: Time averaged mean velocity; Vmax: Maximum velocity; HA: Hepatic artery; HV: Hepatic vein; RI: Resistive artery; BF: Blood flow.

The ROC plots demonstrate that results for liver echogenicity were consistent but had poor diagnostic accuracy[27,28], results for hepatic vein pulsatility were highly variable[18,29,30], results for liver parenchyma[17,23,31], portal vein maximum velocity[23,30,32], and spleen size[16,23,30] were variable, results for caudate to right lobe ratio were consistent but fair in diagnostic accuracy, and results for liver surface consistently had moderate diagnostic accuracy[18,19,23,31,33,34] except for two outlying reports[26,35].

Reference test blinding (QUADAS item 11) was the only item of methodology quality which demonstrated an obvious trend when plotted on a ROC for diagnostic accuracy; most studies which clearly reported blinding performed better than the other studies (Figure 1C).

ROC plots of diagnostic accuracy across disease characteristics (histology staging definition, prevalence, disease aetiology and severity of disease) demonstrated no obvious patterns except that diagnostic accuracy was generally lower for populations with lower prevalence of disease (Figure 2).

DISCUSSION

The aim of this review was to assess the results and quality of studies reporting the diagnostic accuracy of ultrasound imaging techniques used to identify patients with CLD in a high risk population. The search was restricted to techniques that used ultrasound imaging techniques. Transient elastography, which has demonstrated good diagnostic performance[36] and is becoming more widely used in hepatology practice, was not included because it is a non-imaging technique and currently is not an option on standard ultrasound equipment. A review to establish the performance of stand alone ultrasound is useful because ultrasound scans are often provided by medical imaging departments that do not have access to elastography.

The search strategy was optimized for sensitivity rather than precision, as recommended by the Cochrane Collaboration[37] with no filters used which could potentially restrict the search. Efforts to identify as many relevant studies as possible included expanding the search to databases beyond MEDLINE and EMBASE, reading the abstracts of all identified studies and "pearling" of reference lists. Pearling was particular valuable with an additional eight studies identified, however, it is possible that relevant studies may have been missed because the search strategy did not include the grey literature and was restricted to English. Across the studies in this review there was a wide
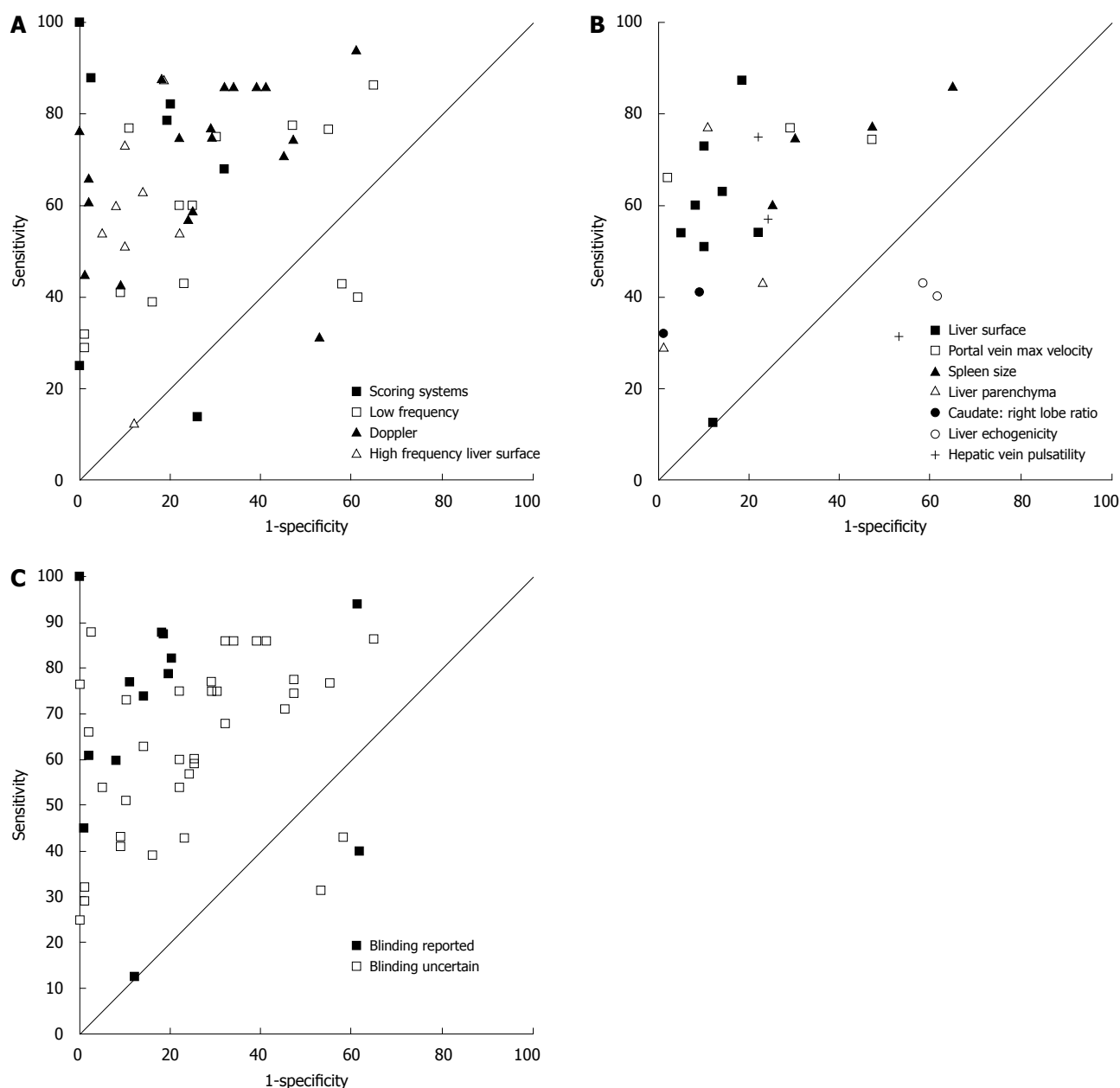
Figure 1 Receiver operator characteristic plot. Diagnostic performance data for categories of ultrasound techniques (A), ultrasound techniques reported multiple times (B) and relating to reference test blinding (C).

range of complexity and clarity of the described ultrasound techniques.

Methodology quality of the included studies was assessed with the QUADAS quality assessment tool, an independently validated method recommended by the Cochrane Collaboration[37]. As recommended[9] the QUADAS tool was modified for the specific needs of the review. Inter-rater variability testing of QUADAS showed good agreement over most of the QUADAS items with nine of 14 having substantial or almost perfect agreement. At the consensus meeting addressing differences in QUADAS ratings it was found that differences tended to relate to differing interpretations of item guidelines. Involving both reviewers in the formulation of the guidelines may have resulted in clearer guidelines and more consistent interpretations.

There was no identifiable group of studies that were clearly superior to the rest nor was there a group of studies that was markedly inferior; therefore all studies in the review were assessed for diagnostic accuracy. Blinding was the only item of methodology quality which demonstrated a relationship with diagnostic accuracy results. Studies reporting blinding for the reference test also reported higher diagnostic accuracy than studies which did not report reference test blinding. This finding further endorses the studies reporting higher diagnostic accuracy, because the chance of bias in these reports is reduced.

The only study characteristic that showed a relationship to diagnostic accuracy was prevalence, with studies reporting low prevalence also tending to have lower diagnostic accuracy. Whilst this may seem surprising, as sensitivity and specificity should be independent of prev-
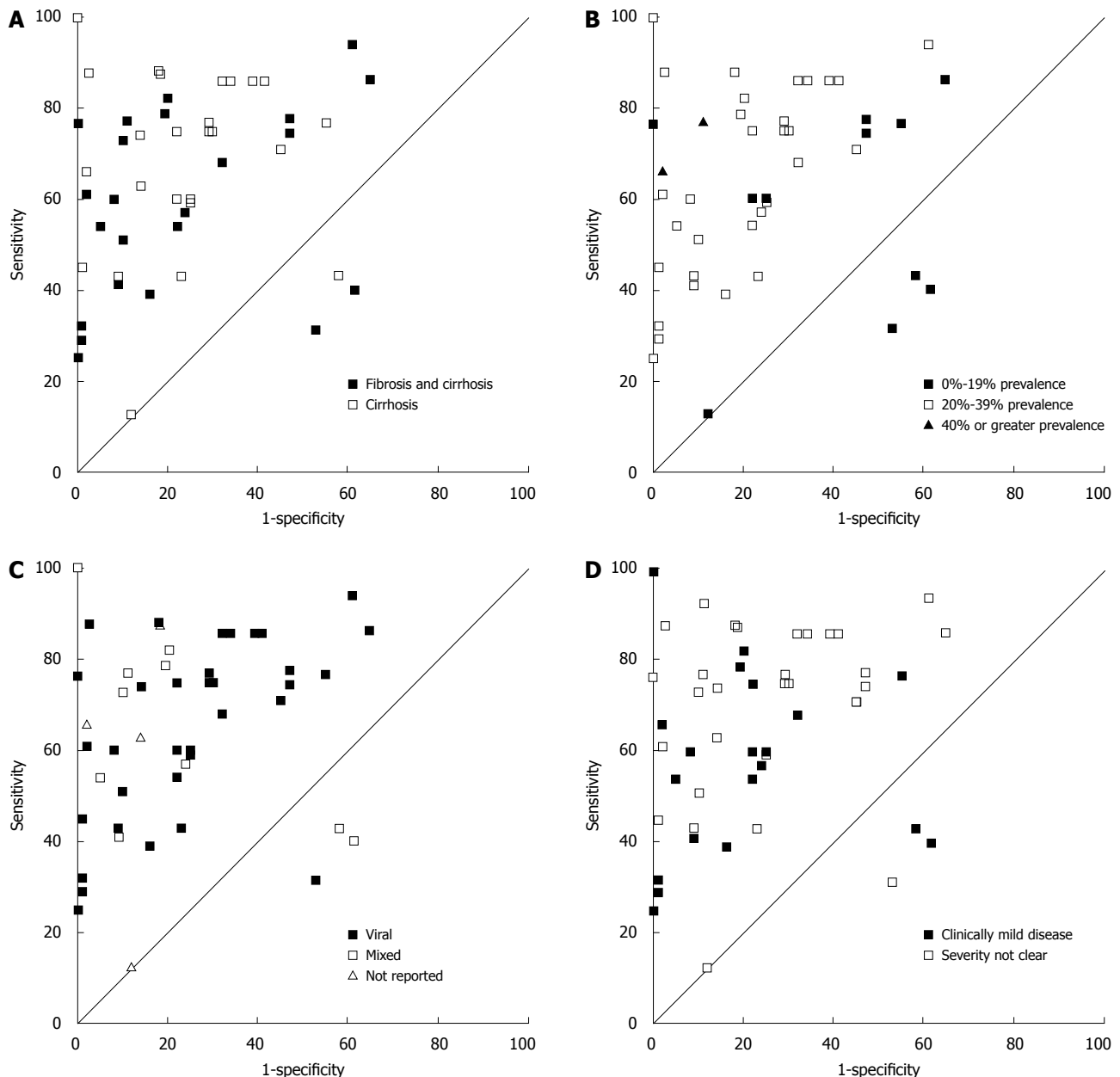
Figure 2 Receiver operator characteristic plot displaying diagnostic accuracy across disease characteristics. A: Histology staging definition; B: Prevalence; C: Disease aetiology; D: Severity of disease.

alence, it has recently been shown that prevalence can affect diagnostic accuracy due to clinical or artefactual variability in studies[38].

Liver biopsy was chosen as the reference test in this review although it has a significant false negative rate due to difficulties with the biopsy technique and sampling error which make it a less than ideal reference test. We justify our choice because it is the test used in clinical practice and is the only practical choice for a reference test. Whilst laparoscopy may be more accurate, it is much more invasive, with significantly more risk, and generally not used in normal clinical practice. Studies using laparoscopy as the reference test were excluded as including more than one reference test has the potential to introduce differential verification bias[9].

Studies were included if the diagnostic accuracy results were either given as true positive (TP), false positive (FP), true negative (TN) and false negative (FN) data or simply in the form of sensitivity and specificity. Restricting studies to those that expressed results in full (TP, FP, TN, FN) would have reduced the range of studies included. Whilst potentially this would have enabled the use of forest plots and meta-analysis to assess the diagnostic accuracy, this was not performed because the numbers of studies of techniques similar enough to enable comparison was too small to provide meaningful results. Instead all studies included in this review were analysed visually using the ROC plot technique. This provided an effective method for comparing data and exploring the relationship between diagnostic accuracy and the quality and characteristics of

the studies[10]. The area under the ROC for the various ultrasound techniques was not calculated due to the lack of reported raw data to make this possible.

Across all studies there was wide variation in both the ultrasound techniques used and in the reported diagnostic sensitivities and specificities for liver fibrosis and cirrhosis. For ultrasound to be clinically useful as a test that can reduce the number of patients requiring liver biopsy it needs to accurately confirm chronic liver disease. To be effective it should have a low false positive rate resulting in high specificity and a high positive predictive value. In this way patients with positive ultrasound results may be able to avoid the risks of liver biopsy. Two studies[22,25] stand out as having very high specificity (100% and 97.6%, respectively) and very high sensitivity (100% and 87.8%, respectively). Both of these studies used scoring systems and this suggests that this may be the best method of identifying severe fibrosis and cirrhosis; however, these results need to be treated with caution. The scoring systems used in both studies were complex, subjective and relied on the compounding of several ultrasound techniques. The use of multiple techniques[20,22,25,31,39] raises concerns regarding reproducibility, as variations may occur with each of the methods used and become magnified with compounding of methods. It is also a concern that in one of these studies[22] it was unclear if blinding had been used, if there were any subject withdrawals, how the selection criteria were applied, how the reference test was applied and how the scoring system was applied. In contrast, the other study[25] scored very well for methodological quality excepting that observer agreement was not reported.

The reporting of observer agreement was poor in many of the reviewed studies despite it being an important consideration when assessing the usefulness of a diagnostic test. We made an assessment of consistency of results across studies which reported similar techniques as a proxy method to determine the reproducibility of a technique in the absence of agreement reporting. Confidence in the results of a study's results can be increased if the technique has been reported over multiple studies with consistent results. We could make this assessment for the following ultrasound techniques; liver echogenicity, caudate lobe to right lobe ratio, portal vein maximum velocity, hepatic vein pulsatility, liver parenchyma echopattern, spleen size and liver surface.

The results for portal vein maximum velocity, hepatic vein pulsatility, liver parenchyma echo-pattern and spleen size were inconsistent between studies.

Consistently poor results of diagnostic accuracy were demonstrated between the two studies which tested measurements of liver echogenicity[27,28]. Liver echogenicity is known to be associated with liver steatosis but not with fibrosis[40] so this result is not surprising. Consistent results of diagnostic accuracy were demonstrated for the caudate lobe to right lobe ratio across two studies[18,31] with high specificity (> 90%) and low sensitivity (41% and 32%, respectively). The liver surface technique was the most frequently reported technique (*n* = 8 reports). Diagnostic accuracy was consistent across six of these studies, with high specificities (78%-95%) and moderate sensitivities (51%-73%)[18,19,23,30,32,34]. These studies were also of reasonable or good methodological quality. There were two studies reporting the liver surface technique[26,34] which produced results that were outliers compared to the other six and contained methodological flaws that were serious enough to not accept their findings. The flaws included an unclear description of patient spectrum or selection criteria in one study[26] together with a reported low prevalence of CLD which does not represent a high risk population which was the population of interest in this review. The other study[35] scored poorly for verification and differential bias and had a significant number of unexplained withdrawals.

The findings of consistent results of diagnostic studies that are methodologically sound make the assessment of liver surface appealing to apply in the clinical environment. This technique also appeared simple to implement, was defined clearly in the reports, and used a simple dichotomous categorical classification technique to interpret definitions of normal and abnormal. Three of these studies[18,19,23] also reported substantial inter and/or intra-observer agreement. Although these studies did not demonstrate high sensitivities, the high specificity and therefore high positive predictive value indicate this technique should be accurate for identifying patients who have a high likelihood of severe fibrosis or cirrhosis and who may benefit by avoiding the risks associated with liver biopsy.

In conclusion, a wide range of ultrasound techniques have been reported in the literature and investigated for their diagnostic accuracy to identify CLD in a high risk population. The most robust ultrasound technique for assessment of CLD appears to be the assessment of liver surface. The studies investigating the liver surface technique consistently demonstrated good observer agreement and high specificity. This review has revealed that an assessment of the liver surface is a useful screen for patients at risk of CLD to assist in determining who should undergo a liver biopsy.

COMMENTS

Background

Chronic liver disease (CLD) is a significant cause of morbidity and mortality. Accurate diagnostic testing to identify early CLD in asymptomatic patients at high risk is advantageous due to recent management and treatment advances. Biopsy, which is the current method of choice, is invasive and carries a significant risk. Less invasive techniques have the potential to reduce biopsy numbers. Ultrasound is one such technique which is readily available, inexpensive and well-tolerated. However, there are several ultrasound techniques in current practice. For an ultrasound study to be clinically useful it has to demonstrate accuracy in confirming CLD. This systematic review informs clinicians of the usefulness of ultrasound in early diagnosis of CLD in high risk patients, in particular, which method is shown to be the most specific and sensitive.

Research frontiers

There have been no identified published systematic reviews addressing diagnostic accuracy in ultrasound of CLD.

Innovations and breakthroughs

This rigorous systematic review identifies methodological and/or reporting flaws in several of the selected papers. It also highlights the variety and range of

diagnostic ultrasound techniques for liver examination in CLD in current usage. This review demonstrates that the most robust ultrasound technique for assessment of CLD appears to be high frequency ultrasound assessment of the liver surface.

Applications

The high specificity of ultrasound of the liver surface provides a clinician with confidence that if signs of CLD are evident then the condition is present. The moderate sensitivity means that if ultrasound signs of CLD are not present, a liver biopsy may be performed to confirm the presence of CLD. Performing high frequency ultrasound of the liver surface in high risk patients has the potential to reduce the number of biopsies in patients at high risk of CLD.

Terminology

Pulse-wave Doppler: A technique by which the ultrasound machine can determine the velocity of blood flowing in vessels. In addition, it allows evaluation of the direction and character of the blood flow. Pulse-wave Doppler is displayed as a spectral waveform on the screen. Maximum velocity: The velocity of blood cells flowing along a vessel will vary according to the position within the blood vessel. The maximum velocity is the greatest velocity detected in a particular vessel in a selected area; pulsatility and resistance indices and the spectral waveform allows quantification of the pulsatility of the blood flow by calculations using the maximum, minimum and mean velocities displayed. The indices are an indication of resistance to blood flow in the vessel and variation from normal may be an indication of disease, either in the vessel itself or the organ it supplies.

Peer review

This is a well written review on the quality and accuracy of ultrasound imaging techniques for identifying patients with chronic liver disease.

REFERENCES

1 **Schuppan D**, Afdhal NH. Liver cirrhosis. *Lancet* 2008; **371**: 838-851

2 **Manning DS**, Afdhal NH. Diagnosis and quantitation of fibrosis. *Gastroenterology* 2008; **134**: 1670-1681

3 **Afdhal NH**, Nunes D. Evaluation of liver fibrosis: a concise review. *Am J Gastroenterol* 2004; **99**: 1160-1174

4 **Brunt EM**. Grading and staging the histopathological lesions of chronic hepatitis: the Knodell histology activity index and beyond. *Hepatology* 2000; **31**: 241-246

5 **Di Lelio A**, Cestari C, Lomazzi A, Beretta L. Cirrhosis: diagnosis with sonographic study of the liver surface. *Radiology* 1989; **172**: 389-392

6 **Gosink BB**, Lemon SK, Scheible W, Leopold GR. Accuracy of ultrasonography in diagnosis of hepatocellular disease. *AJR Am J Roentgenol* 1979; **133**: 19-23

7 **Ohta M**, Hashizume M, Tomikawa M, Ueno K, Tanoue K, Sugimachi K. Analysis of hepatic vein waveform by Doppler ultrasonography in 100 patients with portal hypertension. *Am J Gastroenterol* 1994; **89**: 170-175

8 **NHMRC**. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines: Stage 2 consultation, National Health and Medical Research Council 2008; viewed 3 August 2008. Available from: URL: http://www.nhmrc.gov.au/guidelines/_files/Stage%202%20Consultation%20Levels%20and%20Grades.pdf

9 **Whiting P**, Rutjes AW, Reitsma JB, Bossuyt PM, 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

10 **Whiting PF**, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ. Graphical presentation of diagnostic information. *BMC Med Res Methodol* 2008; **8**: 20

11 **Bedossa P**, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. *Hepatology* 1996; **24**: 289-293

12 **Ishak K**, Baptista A, Bianchi L, Callea F, De Groote J, Gudat F, Denk H, Desmet V, Korb G, MacSween RN. Histological grading and staging of chronic hepatitis. *J Hepatol* 1995; **22**: 696-699

13 **Desmet VJ**, Gerber M, Hoofnagle JH, Manns M, Scheuer PJ. Classification of chronic hepatitis: diagnosis, grading and staging. *Hepatology* 1994; **19**: 1513-1520

14 **Knodell RG**, Ishak KG, Black WC, Chen TS, Craig R, Kaplowitz N, Kiernan TW, Wollman J. Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis. *Hepatology* 1981; **1**: 431-435

15 **Scheuer PJ**. Classification of chronic viral hepatitis: a need for reassessment. *J Hepatol* 1991; **13**: 372-374

16 **Shen L**, Li JQ, Zeng MD, Lu LG, Fan ST, Bao H. Correlation between ultrasonographic and pathologic diagnosis of liver fibrosis due to chronic virus hepatitis. *World J Gastroenterol* 2006; **12**: 1292-1295

17 **Joseph AE**, Saverymuttu SH, al-Sam S, Cook MG, Maxwell JD. Comparison of liver histology with ultrasonography in assessing diffuse parenchymal liver disease. *Clin Radiol* 1991; **43**: 26-31

18 **Colli A**, Fraquelli M, Andreoletti M, Marino B, Zuccoli E, Conte D. Severe liver fibrosis or cirrhosis: accuracy of US for detection--analysis of 300 cases. *Radiology* 2003; **227**: 89-94

19 **Colli A**, Colucci A, Paggi S, Fraquelli M, Massironi S, Andreoletti M, Michela V, Conte D. Accuracy of a predictive model for severe hepatic fibrosis or cirrhosis in chronic hepatitis C. *World J Gastroenterol* 2005; **11**: 7318-7322

20 **Gaiani S**, Gramantieri L, Venturoli N, Piscaglia F, Siringo S, D'Errico A, Zironi G, Grigioni W, Bolondi L. What is the criterion for differentiating chronic hepatitis from compensated cirrhosis? A prospective study comparing ultrasonography and percutaneous liver biopsy. *J Hepatol* 1997; **27**: 979-985

21 **Liu CH**, Hsu SJ, Lin JW, Hwang JJ, Liu CJ, Yang PM, Lai MY, Chen PJ, Chen JH, Kao JH, Chen DS. Noninvasive diagnosis of hepatic fibrosis in patients with chronic hepatitis C by splenic Doppler impedance index. *Clin Gastroenterol Hepatol* 2007; **5**: 1199-1206.e1

22 **Xu Y**, Wang B, Cao H. An ultrasound scoring system for the diagnosis of liver fibrosis and cirrhosis. *Chin Med J* (Engl) 1999; **112**: 1125-1128

23 **Iliopoulos P**, Vlychou M, Margaritis V, Tsamis I, Tepetes K, Petsas T, Karatza C. Gray and color Doppler ultrasonography in differentiation between chronic viral hepatitis and compensated early stage cirrhosis. *J Gastrointestin Liver Dis* 2007; **16**: 279-286

24 **Paggi S**, Colli A, Fraquelli M, Viganò M, Del Poggio P, Facciotto C, Colombo M, Ronchi G, Conte D. A non-invasive algorithm accurately predicts advanced fibrosis in hepatitis C: a comparison using histology with internal-external validation. *J Hepatol* 2008; **49**: 564-571

25 **Nishiura T**, Watanabe H, Ito M, Matsuoka Y, Yano K, Daikoku M, Yatsuhashi H, Dohmen K, Ishibashi H. Ultrasound evaluation of the fibrosis stage in chronic liver disease by the simultaneous use of low and high frequency probes. *Br J Radiol* 2005; **78**: 189-197

26 **Ladenheim JA**, Luba DG, Yao F, Gregory PB, Jeffrey RB, Garcia G. Limitations of liver surface US in the diagnosis of cirrhosis. *Radiology* 1992; **185**: 21-23; discussion 23-24

27 **Mathiesen UL**, Franzén LE, Aselius H, Resjö M, Jacobsson L, Foberg U, Frydén A, Bodemar G. Increased liver echogenicity at ultrasound examination reflects degree of steatosis but not of fibrosis in asymptomatic patients with mild/moderate abnormalities of liver transaminases. *Dig Liver Dis* 2002; **34**: 516-522

28 **Hultcrantz R**, Gabrielsson N. Patients with persistent elevation of aminotransferases: investigation with ultrasonography, radionuclide imaging and liver biopsy. *J Intern Med* 1993; **233**: 7-12

29 **Colli A**, Cocciolo M, Riva C, Martinez E, Prisco A, Pirola M, Bratina G. Abnormalities of Doppler waveform of the hepatic veins in patients with chronic liver disease: correlation with histologic findings. *AJR Am J Roentgenol* 1994; **162**: 833-837

30 **Schneider AR**, Teuber G, Kriener S, Caspary WF. Nonin-

vasive assessment of liver steatosis, fibrosis and inflammation in chronic hepatitis C virus infection. *Liver Int* 2005; **25**: 1150-1155

31   **D'Onofrio M**, Martone E, Brunelli S, Faccioli N, Zamboni G, Zagni I, Fattovich G, Pozzi Mucelli R. Accuracy of ultrasound in the detection of liver fibrosis in chronic viral hepatitis. *Radiol Med* 2005; **110**: 341-348

32   **Cioni G**, D'Alimonte P, Cristani A, Ventura P, Abbati G, Tincani E, Romagnoli R, Ventura E. Duplex-Doppler assessment of cirrhosis in patients with chronic compensated liver disease. *J Gastroenterol Hepatol* 1992; **7**: 382-384

33   **Viganò M**, Visentin S, Aghemo A, Rumi MG, Ronchi G. US features of liver surface nodularity as a predictor of severe fibrosis in chronic hepatitis C. *Radiology* 2005; **234**: 641; author reply 641

34   **Gaia S**, Cocuzza C, Rolle E, Bugianesi E, Carucci P, Vanni E, Evangelista A, Rizzetto M, Brunello F. A comparative study between ultrasound evaluation, liver stiffness and biopsy for staging of hepatic fibrosis in patients with chronic liver disease. *J Hepatol* 2009; **50** (Suppl 1): S361

35   **Ferral H**, Male R, Cardiel M, Munoz L, Quiroz y Ferrari F. Cirrhosis: diagnosis by liver surface analysis with high-frequency ultrasound. *Gastrointest Radiol* 1992; **17**: 74-78

36   **Shaheen AA**, Wan AF, Myers RP. FibroTest and FibroScan for the prediction of hepatitis C-related fibrosis: a systematic review of diagnostic test accuracy. *Am J Gastroenterol* 2007; **102**: 2589-2600

37   **de Vet HCW**, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, Chapter 7: Searching for Studies; 0.4 ed. 2008 The Cochrane Collaboration

38   **Leeflang MM**, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol* 2009; **62**: 5-12

39   **Wang JH**, Changchien CS, Hung CH, Eng HL, Tung WC, Kee KM, Chen CH, Hu TH, Lee CM, Lu SN. FibroScan and ultrasonography in the prediction of hepatic fibrosis in patients with chronic viral hepatitis. *J Gastroenterol* 2009; **44**: 439-446

40   **Saverymuttu SH**, Joseph AE, Maxwell JD. Ultrasound scanning in the detection of hepatic fibrosis and steatosis. *Br Med J* (Clin Res Ed) 1986; **292**: 13-15