

CME

Accuracy of Ultrasonography, Spiral CT, Magnetic Resonance, and Alpha-Fetoprotein in Diagnosing Hepatocellular Carcinoma: A Systematic Review

Agostino Colli, M.D.,¹ Mirella Fraquelli, M.D., Ph.D.,² Giovanni Casazza, Ph.D.,³ Sara Massironi, M.D.,¹ Alice Colucci, M.D.,¹ Dario Conte, M.D.,² and Piergiorgio Duca, M.D.³

¹Department of Internal Medicine, Ospedale "A. Manzoni", Lecco, ²Postgraduate School of Gastroenterology, IRCCS Ospedale Maggiore, Milan, and ³Department of Clinical Sciences, Ospedale "L. Sacco", Milan, Italy

BACKGROUND AND AIM: In patients with chronic liver disease, the accuracy of ultrasound scan (US), spiral computed tomography (CT), magnetic resonance imaging (MRI), and alpha-fetoprotein (AFP) in diagnosing hepatocellular carcinoma (HCC) has never been systematically assessed, and present systematic review was aimed at this issue.

METHODS: Pertinent cross-sectional studies having as a reference standard pathological examinations of the explanted liver or resected segment(s), biopsies of focal lesion(s), and/or a period of follow-up, were identified using MEDLINE, EMBASE, Cochrane Library, and CancerLit. Pooled sensitivity, specificity, and likelihood ratios (LR) were calculated using the random effect model. Summary receiver operating characteristic (SROC) curve and predefined subgroup analyses were made when indicated.

RESULTS: The pooled estimates of the 14 US studies were 60% (95% CI 44–76) for sensitivity, 97% (95% CI 95–98) for specificity, 18 (95% CI 8–37) for LR+, and 0.5 (95% CI 0.4–0.6) for LR–; for the 10 CT studies sensitivity was 68% (95% CI 55–80), specificity 93% (95% CI 89–96), LR+ 6 (95% CI 3–12), and LR– 0.4 (95% CI 0.3–0.6); for the nine MRI studies sensitivity was 81% (95% CI 70–91), specificity 85% (95% CI 77–93), LR+ 3.9 (95% CI 2–7), and LR– 0.3 (95% CI 0.2–0.5). The sensitivity and specificity of AFP varied widely, and this could not be entirely attributed to the threshold effect of the different cutoff levels used.

CONCLUSIONS: US is highly specific but insufficiently sensitive to detect HCC in many cirrhotics or to support an effective surveillance program. The operative characteristics of CT are comparable, whereas MRI is more sensitive. High-quality prospective studies are needed to define the actual diagnostic role of AFP.

(Am J Gastroenterol 2006;101:513–523)

INTRODUCTION

Hepatocellular carcinoma (HCC) is a major medical problem worldwide, particularly in areas with a high prevalence of chronic hepatitis B and/or C viral infections (1, 2). In the vast majority of cases, HCC develops on an underlying cirrhosis, although a few cases have been reported in people without this condition (3). Despite recent improvement, the prognosis of HCC remains very poor as only about 10% of the patients can receive curative treatment as orthotopic liver transplantation (OLT) or surgical resection, which are unfeasible in most cases due to severe clinical deterioration at diagnosis and/or the inaccuracy of preoperative clinical evaluation and staging procedures (4).

HCC is currently diagnosed on the basis of the results of the serological determination of alpha-fetoprotein levels

(AFP), liver ultrasound scans (US), spiral computed tomography (CT), and magnetic resonance imaging (MRI) techniques (5), but their accuracy has never been systematically assessed.

The aim of this study was to review the accuracy of AFP, US, spiral CT, and MRI in diagnosing HCC in patients with chronic liver disease, using as the reference standard the pathological findings of the explanted liver or resected hepatic segment(s), histological specimen(s) obtained from focal lesion(s), or an adequate period of follow-up.

METHODS

Data Sources

Pertinent primary studies, with no language restriction, were retrieved using the MEDLINE (from 1966 to December 2004), EMBASE (from 1988 to December 2004), and Cochrane Library and CancerLit databases. Reference lists

To access a continuing medical education exam for this article, please visit www.acg.gi.org/journalcme.

from all of the available review articles, primary studies, and proceedings of major meetings (from 1995 to December 2004) were also considered. When the data were incomplete, the original material was requested directly from the authors. When the MEDLINE and EMBASE databases were examined, the search strategy included both medical subject headings (MeSH) terms and free language words.

Study Selection Criteria

Cross-sectional studies assessing the HCC diagnostic accuracy of AFP, US, spiral CT, and MRI (alone or in combination, but not sequentially) were considered when their reference standard was the pathology of the explanted or resected liver, or the histology of focal liver lesion(s). A follow-up period of at least 6 months, to allow the confirmation of an initial negative result (usually a normal AFP and/or negative US liver scan), was also regarded as a reference standard in order to avoid a potential work-up bias. This length of follow-up was also chosen as it could add information on potentially synchronous lesions from the parenchyma surrounding the resected or biopsied area.

Studies using sequential test combinations (*e.g.*, US liver scans in patients selected on the basis of AFP concentrations) were excluded because the selection of patients on the basis of diagnostic test results could have unpredictably modified the estimate of the operative characteristics of the tests themselves (6).

Type of Participants

The review only examined studies including patients with chronic liver disease (*i.e.*, cirrhosis or chronic hepatitis) assessed with the aim of detecting the possible presence of HCC.

Type of Index Tests

The diagnostic accuracy of AFP, US, spiral CT, and MRI was tested against one of the above reference standards in patients with chronic liver disease and a focal liver lesion, found during the course of a surveillance program or in a clinical setting.

Quality Assessment of Primary Studies

All of the included studies were assessed for their methodological quality according to previously defined standards (7, 8) on the basis of their *study design* (cohort or case-control), *spectrum composition* (reflecting or not the representativeness of the included patients in relation to those undergoing tests in clinical practice), *patient selection* (consecutive or not), and *verification* (complete or partial: *i.e.*, when all or only some of the patients underwent the reference standard; or different: *i.e.*, when more than one reference standard was used); moreover, the *time interval* between the index test and the reference standard, the *interpretation of the test results* (blinded or not), *data collection* (prospective, retrospective, or unknown), and the *details concerning the test,*

reference standard, or population (sufficient or insufficient) were considered.

Type of Outcome Measures

The sensitivity, specificity, and positive and negative likelihood ratios (LR+ and LR–, respectively) were the outcome measures used to evaluate diagnostic accuracy.

Methods of the Review

All of the studies identified as described above were analyzed by four reviewers (A.C., M.F., A.C., and S.M.), each of whom reexamined them in order to confirm those fulfilling the inclusion criteria, and graded their methodological quality on the basis of previously reported criteria (7, 8). The data concerning the types of participants, interventions, and outcome measures were independently extracted by the reviewers, who openly discussed any discrepancy; only in the case of disagreement was the further and definite judgment of an independent clinical expert (DC) applied.

Statistical Analysis

The sensitivity (*i.e.*, true positive rate, TPR) and specificity (*i.e.*, true negative rate, TNR) of each study were defined, and exact 95% confidence intervals (95% CI) were calculated on the basis of a binomial distribution. Their LR+ and LR– were also defined, and the 95% CI calculated on the basis of a normal approximation of their log-transformation.

Considering the low methodological quality of most of the diagnostic studies available in the literature, it seemed to us advisable to use the more conservative random effect model, in order to pool the estimates of sensitivity and specificity, even if there was no apparent heterogeneity (9, 10).

The homogeneity of the TPR and TNR estimates was evaluated using the Fisher's exact test (11), with a level of significance of 0.1 (12). In the case of heterogeneity, two further steps were taken (13). First, a summary receiver operating characteristic (SROC) analysis (14) was made only after having demonstrated the possible presence of a cutoff effect by calculating Spearman's correlation coefficient ρ between TPR and FPR (the cutoff effect was considered present in the case of a ρ value >0.4) (10) and, second, the predefined subgroup analysis was also performed.

A subgroup analysis was planned for predefined subsets of studies in which the pathology of the explanted liver represented the reference standard, the prevalence of cirrhosis and HBV infection was $>75\%$ and $\geq 15\%$, respectively (thus supporting an endemic infection), the frequency of HCC was $\geq 30\%$ (thus suggesting the selection of the tested population), only symptomatic patients had been included (*i.e.*, those with weight loss, fever, jaundice, and/or rapidly deteriorating liver function, all of which are consistent with a possible underlying HCC). The subgroup analysis also included US, CT, and MRI studies carried out after 1985, 1997, and 2000, respectively.

In order to check for a possible publication bias, a funnel plot of the individual studies was made with logDORs

(logarithm of the diagnostic odds ratios) being plotted against the sample size; an asymmetrical funnel plot suggested that other small studies may have been conducted but not published because of unfavorable results (15).

All of the statistical analyses were made using SAS statistical software version 8.2 (SAS Institute Inc., Cary, NC, USA).

RESULTS

The study selection process is detailed in Figure 1. Given the 40% overlap between the databases, 2,524 of the 4,207 primary studies were retrieved in abstract form, and the full text was obtained of the 88 consistent with the aim of the review. Fifty-eight of these studies were excluded because of the poor definition of the investigated test or reference standard (14 studies), the inappropriate use of the reference standard (14 studies), the absence of sensitivity and/or specificity estimates (29 studies), or the sequential use of tests (1 study). A total of 30 studies (29 in English and 1 in German) (16–45) fulfilled all of the inclusion criteria and were considered for the analysis (see Table 1).

US

The sensitivity and specificity of the 14 selected US studies (16–19, 21, 23, 26, 27, 30, 31, 35, 40–42) ranged from 30% to 100% and from 73% to 100%, respectively; the pooled estimates and the corresponding LR+ and LR– are shown in Table 2. The heterogeneity of the estimates was demonstrated by means of a formal statistical test (exact test: $p < 0.001$ for both sensitivity and specificity). As the ρ value was less than 0.4, the SROC curve was not calculated (Fig. 2) but a subgroup analysis was carried out.

The sensitivity and specificity estimates of the eight studies, using the histological findings of the explanted liver as the reference standard (21, 23, 27, 30, 31, 35, 40, 42) were still heterogeneous, with the following pooled values: sensitivity 48% (95% CI 34–62%), specificity 97% (95% CI 95–98%), LR+ 8.2 (95% CI 5.3–12.7), and LR– 0.6 (95% CI 0.5–0.7).

It was possible to obtain and SROC curve for this subgroup (see Fig. 3).

The further predefined subgroup analyses were inconclusive or precluded by the paucity of the studies. In detail, one study carried out before 1985 involved only symptomatic patients with a prevalence of HCC and HBV of 36% and 80%, respectively (17); two studies did not indicate the actual prevalence of cirrhosis (19, 27), and further two had a prevalence of HCC of >30% (30, 31). Furthermore, most of the studies lacked data concerning the prevalence of chronic HBV and/or HCV infection (Table 1) and, finally, the time interval between US and the reference standard was rarely reported or, when reported, had an extremely large range.

AFP

The sensitivity, specificity, LR+, and LR– of the nine studies (20, 22, 25–27, 33, 41, 44, 45) assessing the diagnostic accuracy of AFP are shown in Table 2. These studies used different cutoff values and the pooled estimates based on the cutoff values are shown in Figure 4. The ρ value of 0.79 allowed to calculate the SROC curve (Fig. 5), which showed that the operative characteristics for the same cutoff value were different.

Spiral CT

The sensitivity, specificity, LR+, and LR– of the 10 studies assessing the diagnostic accuracy of spiral CT (26–30, 32, 36, 37, 40, 42) and their pooled estimates are shown in Table 3. A formal statistical test (exact test: $p < 0.001$ for both sensitivity and specificity) demonstrated the heterogeneity of the results, despite the fact that all of the studies had involved OLT candidates and used the pathology of the explanted liver as the reference standard. The ρ value of <0.4 (Fig. 5) prevented the calculation of an SROC curve, and the further subgroup analyses were inconclusive or precluded by the paucity of the studies (see Table 2).

MRI

The operative characteristics of MRI and the pooled estimates of nine pertinent studies (24, 30, 34, 36, 38–40, 42, 43) are shown in Table 3. Once again, a formal statistical test (exact test: $p < 0.001$ for both sensitivity and specificity) demonstrated the heterogeneity of the results, although all of the studies had involved OLT candidates and used the pathology of the explanted liver as the reference standard. The calculation of an SROC curve was prevented by the ρ values of < 0.4 (Fig. 2), and the further subgroup analyses were inconclusive or precluded by the paucity of the studies.

In order to assess a possible publication bias, scatter plots were designed of the logDORs of the individual studies against their sample size. The funnel plots for AFP, US, CT, and MRI are given in Figure 6: in details, the CT and MRI plots showed wide scattering (without any evidence of clear asymmetry), the AFP plot showed marked asymmetry (with small studies missing from the bottom left quadrant, and thus suggesting a publication bias or other small

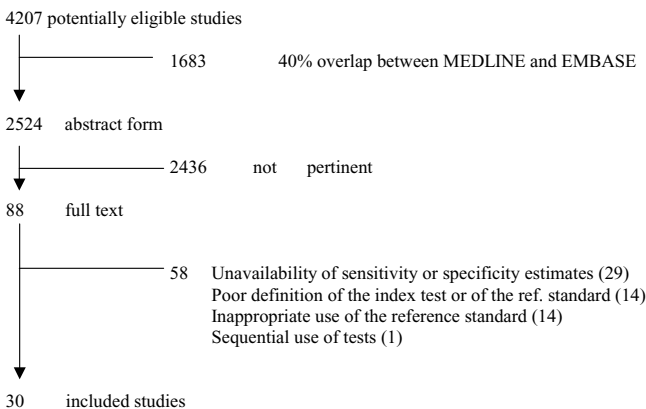


Figure 1. Article selection process.

Table 1. Main Characteristics of the Included Studies

Author	Yr	Study Type	Symptoms*	Index Test	Time Interval (days)	Reference Standard	Cirrhotics (%)	HCC Rate (n/n)	HBV Rate (n/n)	HCV Rate (n/n)
Okazaki	84	PC	No	US	NA	Hist F-U	93	14/245	44/245	NA
Maringhini	84	PC	Yes	US	NA	Hist	100	24/67	54/67	NA
Kobayashi	85	PC	No	US	NA	Hist F-U	100	8/95	33/95	NA
Tanaka	86	PC	No	US	NA	HCC reg F-U	NA	113/5,339	NA	NA
Piantino	89	CC	–	AFP	NA	Surgery	97 (all HCC)	333/766	90/333	NA
”		PC	No	AFP	NA	F-U	100	21/209	NA	NA
Dodd	92	PC	No	US	1–343	OLT	100	28/200	NA	NA
Lopez	96	CC	NA	AFP	NA	Hist	NA (cases) 42 (CRL)	80/62+76	NA	NA
Saada	97	PC	NA	US	NA	OLT	100	6/39	NA	NA
Born	98	RC	NA	MRI	20–94	”	100	19/47	NA	NA
Bayati	98	CC	–	AFP	NA	”	54	15/200	0/200	200/200
Chalsani	99	RC	No	US	NA	Hist	100	27/285	NA	35/285
”			”	S-CT	NA	”	100	”	NA	”
”			”	AFP	NA	”	100	”	NA	”
Gambarin	00	PC	No	US	<180	OLT	NA	19/106	6/106	41/106
”	”	PC	”	S-CT	”	”	”	”	”	”
”			”	AFP	”	”	”	”	”	”
Lim	00	PC	No	S-CT	50–100	OLT	100	15/41	41/41	0/41
Peterson	00	PC	No	S-CT	1–671	”	100	44/320	16/320	106/320
”		RC	”	S-CT	”	”	”	44/320	”	”
Rode	01	PC	No	US	NA	OLT	100	13/43	4/43	19/43
”			”	S-CT	”	”	”	”	”	”
”			”	MRI	”	”	”	”	”	”
Kim	00	RC	No	US	NA	OLT	100	16/52	49	NA
Mortelè	01	RC	”	S-CT	”	OLT	”	17/53	”	”
Trevisani	01	CC	–	AFP	NA	Hist F-U	92 (all HCC)	170/340	26/170	120/170
Krinsky	01	PC	No	MRI	NA	OLT	100	11/171*	10/171	28/171
Bennet	02	CR	No	US	<90	OLT	100	27/200	8/200	70/200
De Ledinghen	02	PC	Yes	S-CT	1–161	OLT	100	21/34	4/34	12/34
”			”	MRI	”	”	”	”	”	”
Zacherl	02	PC	Yes	S-CT	<1	OLT	100	23/23	4/23	7
Bhartia	02	PC	No	MRI	3–245	OLT	100	14/31	5	17
Mori	02	PC	No	MRI	NA	OLT	98	20/48	8	19
Teefey	03	PC	No	MRI	30–450	OLT	100	9/25	NA	NA
”		”	”	CT	”	”	”	”	”	”
”		”	”	US	”	”	”	”	”	”
Tong	01	PC	No	AFP	NA	Hist F-U	29	31/602	163	439
”		”	”	US	”	”	”	”	”	”
Libbrecht	03	PC	”	US	”	OLT	100	17/49	10/49	13/49
”	03	PC	”	CT	”	OLT	100	”	”	”
”	”	”	”	MRI	”	”	”	”	”	”
Burrel	03	PC	Yes	MRI	(Mean 46)	OLT	100	29/50	3	36
Nguyen	02	CC	NA	AFP	0–300	Hist	100	163/312	0	100
Marrero	03	CC	NA	AFP	”	Hist	NA	55/207	12/207	100/207

PC = prospective cohort; RC = retrospective cohort; CC = case control; OLT = orthotopic liver transplantation; NA = not assessed; US = ultrasonography; S-CT = spiral CT; MRI = magnetic resonance; AFP = alpha-fetoprotein; F-U = follow-up; Hist = histology; HCC reg = HCC registry; CRL = controls.

*Population selected on the basis of symptoms suggesting HCC (see text).

study effect), and the US one was also asymmetric (a lack of studies in the bottom right quadrant). This last asymmetry did not seem to be related to a publication bias but to an overestimate of DORs for the larger studies that used follow-up as the reference standard; the small studies using more rigorous reference standards showed less diagnostic accuracy.

Quality Assessment

In terms of the quality assessment (Table 4), the same reference standard was used for all of the patients in 20 studies

(17, 19, 23, 24, 27–40, 42, 43), thus leading to a complete verification and lack of bias; the verification of the remaining 10 (16, 18, 20–22, 25, 26, 41, 44, 45) was partial as different reference standards had been used in different subsets of patients. A significantly higher number of studies using OLT as the reference standard (17/19) were completely verified as against only three of the 11 non-OLT studies. Blinding was surely used in eight studies (30, 36–40, 42, 43), but the remaining 22 studies (16–29, 31–35, 41, 44, 45) were either not blinded or their blinding status was not defined; patient recruitment was consecutive in 17 studies (21, 23–31, 37, 39,

Table 2. Accuracy of US (upper panel) and Alpha-Fetoprotein (lower panel) in Diagnosing HCC

Author	Yr	References	Sensitivity (%)	Specificity (%)	Likelihood Ratio	
					Positive	Negative
Okazaki	84	16	86	99	66	0.14
Maringhini	84	17	92	86	6.5	0.09
Kobayashi	85	18	75	98	32.6	0.26
Tanaka	86	19	47	100	589	0.41
Dodd	92	21	43	98	21.5	0.58
Saada	97	23	33	100	333	0.67
Chalasani	99	26	59	92	8.4	0.45
Gambarin	00	27	58	94	9.6	0.44
Rode	01	30	46	95	9.2	0.57
Kim	01	31	38	92	4.7	0.67
Bennett	01	35	30	97	7.4	0.72
Teefey	03	40	89	73	3.3	0.15
Tong	01	41	100	98	50	0.0
Libbrecht	03	42	40	100	400	0.6
Pooled estimates (95% CI)			60.5 (44–76)	96.9 (95–98)	17.7 (8.5–36.9)	0.5 (0.4–0.6)

Author	Yr	References	Cut-off (ng/mL)	Sensitivity (%)	Specificity (%)	Likelihood Ratio	
						Positive	Negative
Piantino	89	20	> 50*	67	87	5.2	0.38
”	”	”	> 50 [†]	76	67	2.3	0.35
”	”	”	> 100 [†]	62	97	20.6	0.39
Lopez	96	22	> 200	59	97	22.5	0.42
Bayati	98	25	> 10	93	67	2.8	0.1
”	”	”	> 17.8	35	99	35	0.65
Chalasani	99	26	> 20	63	86	4.8	0.42
Gambarin	00	27	> 20	58	91	6.4	0.46
”	”	”	> 50	47	96	11.8	0.6
Trevisani	01	33	> 20	60	91	6.6	0.43
Tong	01	41	> 11	86	89	7.8	0.15
”	”	”	> 21	41	94	6.8	0.62
Nguyen	02	44	> 10	78	61	2.0	0.36
”	”	”	> 20	63	80	3.1	0.46
”	”	”	> 50	51	89	4.6	0.55
”	”	”	> 100	41	97	13.6	0.61
”	”	”	> 200	32	100	320	0.68
Marrero	02	45	> 11	77	71	2.6	0.32
”	”	”	> 20	68	86	4.8	0.37
”	”	”	> 100	47	98	23.5	0.54

*Cross-sectional study.
[†]Longitudinal study.

41–45), nonconsecutive in three (22, 32, 33), and not defined in 10 (16–20, 34–36, 38, 40). All but seven studies (16, 18, 20, 22, 25, 26, 44) detailed the characteristics of both tests under investigation (*i.e.*, the AFP laboratory test and US, CT,

and MRI equipment and procedures) and the reference standards. Only two studies (39, 40) (both of which used OLT as the reference standard) gave data concerning interobserver variability.

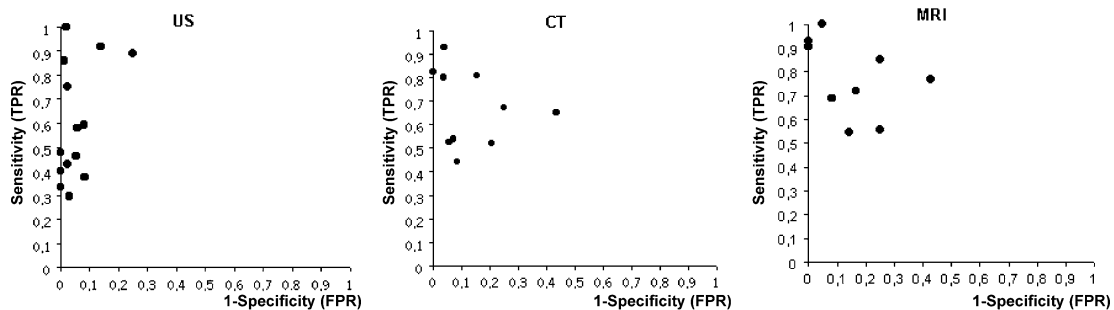


Figure 2. Relationship between sensitivity (TPR) and 1-specificity (FPR) of US, CT, and MRI in diagnosing HCC in the included primary studies (circles). For US, CT, and MRI the ρ value was <0.4 and prevented the calculation of the SROC curve.

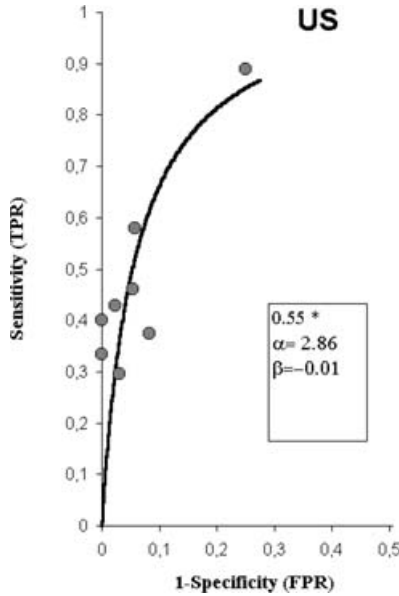


Figure 3. Summary receiver operating characteristic (SROC) curve for the ultrasonographic detection of HCC in the studies having OLT as the reference standard. Asterisk refers to the value of Spearman’s correlation coefficient (ρ).

DISCUSSION

This systematic review evaluated the accuracy of non-invasive diagnostic techniques (*i.e.*, serum AFP concentrations, US, spiral CT, and MRI) in detecting HCC. In order to minimize the risk of overlooking pertinent primary studies, a very sensitive search strategy was chosen due to the lack of defined guidelines for the retrieval of studies in the field of diagnostic test (10). Thus, only 88 out of 2,524 retrieved studies were consistent with the aim of the review and only 30 (1%) of them satisfied the predefined inclusion criteria, thus allowing the calculation of their operative characteristics (*i.e.*, sensitivity and specificity) and their systematic analysis (Fig. 1).

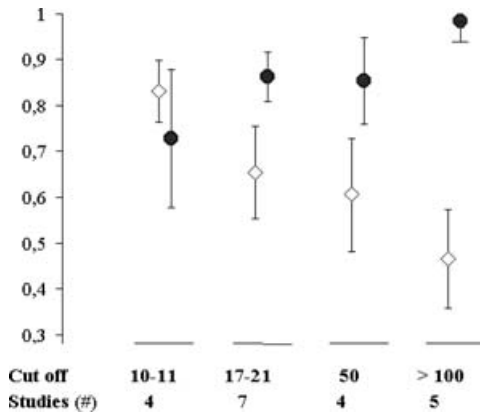


Figure 4. Pooled estimates of sensitivity and specificity (with 95% confidence intervals) of AFP by the different cut-off ranges obtained from different studies. Open diamonds refer to sensitivity and close circles to specificity.

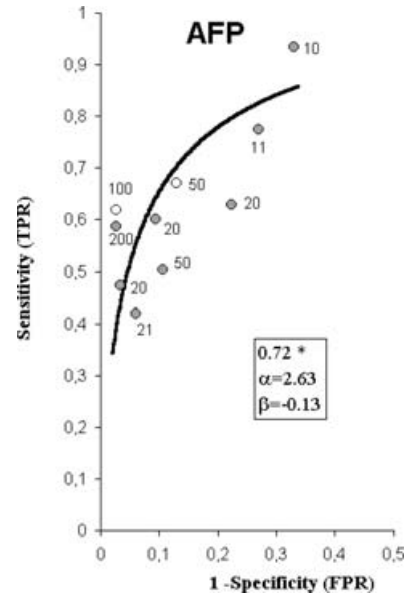


Figure 5. Summary receiver operating characteristic (SROC) curve for the detection of HCC by alpha-fetoprotein. The figures near circles indicate the cutoff values used in the different studies. The two open circles indicate the two different population with different AFP cut-offs (50 and 100 ng/mL, respectively) examined in the Piantino’s study (20). Asterisk refers to Sperman’s correlation coefficient.

As expected, the largest number of studies used US as the diagnostic test. Overall, the 14 studies consistent with the aim of the review showed a wide range of sensitivity (30–100%) and specificity (73–100%), which could reflect differences in the operator skills and experience. Other possible explanations for this wide variation may be due to differences in the tested populations, different indications for performing the test and/or differences in the stage of liver disease. It is known that population selection seems to affect the operative characteristic of diagnostic tests in an unpredictable manner (6), for example, in a selected population of HBsAg chronic carriers with high AFP levels (46), US was more sensitive (86%) and less specific (82%) in diagnosing HCC than reported in the present review. However, our predefined subgroup analyses for many known factors failed to demonstrate the supposed differences in the tested populations, mainly because of a lack of pertinent data. In detail, the high or low prevalence of HCC, the presence of cirrhosis *versus* chronic hepatitis, the presence or absence of symptoms, or the high *versus* low prevalence of HBsAg or anti-HCV positivity did not allow to explain this heterogeneity. Moreover, differences in tumor size may also have been responsible because large HCCs are obviously more easily detectable, and the definition of the minimum detectable diameter of a given focal liver lesion can be greatly affected by the technical performances of the US equipment. The best subgroup for exploring the relevance of HCC size may be that of the OLT patients, as it is more homogeneous. However, although it was possible to design an SROC curve suggesting a threshold effect, the lack of precise data (only mean and/or range values were available)

Table 3. Accuracy of Spiral CT (upper panel) and MRI (lower panel) in Diagnosing HCC

Author	Yr	References	Sensitivity (%)	Specificity (%)	Likelihood Ratio	
					Positive	Negative
Chalasanı	99	26	93	96	22.7	0.09
Gambarın	00	27	53	94	8.8	0.5
Lim	00	28	80	96	20	0.21
Peterson	00	29	44	92	5.1	0.72
Rode	01	30	54	93	7.7	0.49
Mortelè	01	32	82	100	82	0.18
De Ledinghen	02	36	81	85	5.2	0.22
Zacherl	01	37	65	56	1.49	0.62
Teefey	03	40	67	72	2.4	0.46
Libbrecht	03	42	50	79	2.4	0.63
Pooled estimates (95% CI)			67.5 (55–80)	92.5 (89–96)	6.1 (3.1–12)	0.4 (0.3–0.6)
Born	98	24	69	92	8.25	0.34
Rode	01	30	77	57	1.79	0.4
Krinsky	01	34	54	86	3.8	0.53
de Ledinghen	02	36	90	100	905	0.09
Mori	02	39	85	74	3.2	0.2
Teefey	03	40	56	72	2.1	0.61
Bhartia	02	38	93	100	900	0.07
Libbrecht	03	42	70	82	0.71	0.37
Burrell	03	43	100	95	20.0	0.00
Pooled estimates (95% CI)			80.6 (70–91)	84.8 (77–93)	3.9 (2.4–6.5)	0.3 (0.2–0.5)

precluded an analysis of the role of HCC size on US operative characteristics and, furthermore, the explicit definition of the minimum detectable diameter was lacking. Finally, subgroup analysis failed to show any difference between the studies

carried out before and after 1985, and thus failed to support the possible role of technological advances in US equipment. In particular, studies assessing the role of color or power-Doppler US (47), or US contrast media (48), which claimed

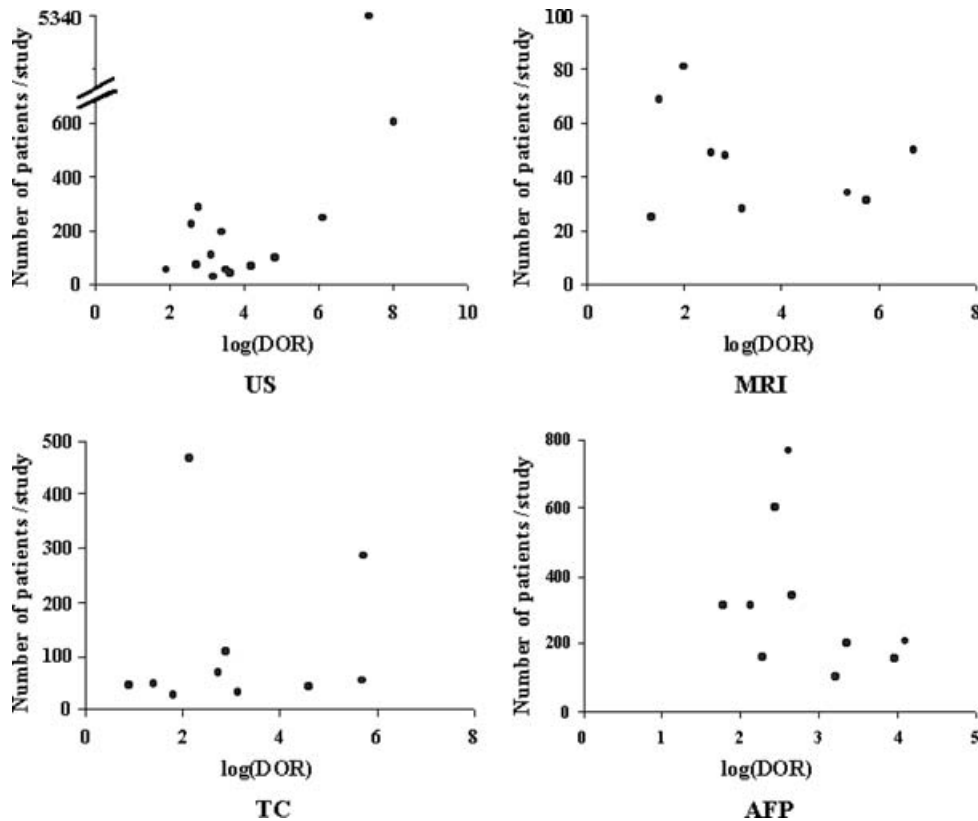


Figure 6. Inverted funnel plot of the individual studies using the four techniques, with the logDORs (diagnostic odds ratio) plotted against sample size (*i.e.*, number of patients/study).

Table 4. Quality of the 30 Studies Evaluating the Diagnostic Performances of AFP, US, CT, and MRI in Diagnosing HCC

Author	Yr	References	Study Type	Verification	Blinding	Consecutive Patients	Test Under Investigation	Ref. Test	Intra/Inter-Observer Variability
Okazaki	84	16	PC	Partial	ND	ND	Yes	No	ND
Maringhini	84	17	PC	Complete	ND	ND	Yes	Yes	ND
Kobayashi	85	18	PC	Partial	ND	ND	No	Yes	ND
Tanaka	86	19	PC	Complete	ND	ND	Yes	Yes	ND
Piantino	89	20	CC	Partial	ND	ND	Yes	No	ND
Dodd	92	21	PC	Partial	ND	Yes	Yes	Yes	ND
Lopez	96	22	CC	Partial	ND	No	Yes	No	ND
Saada	97	23	PC	Complete	ND	Yes	Yes	Yes	ND
Born	98	24	RC	Complete	ND	Yes	Yes	Yes	ND
Bayati	98	25	CC	Partial	ND	Yes	No	No	ND
Chalasani	99	26	RC	Partial	ND	Yes	No	No	ND
Gambarin	00	27	PC	Complete	ND	Yes	Yes	Yes	ND
Lim	00	28	PC	Complete	ND	Yes	Yes	Yes	ND
Peterson	00	29	PC	Complete	ND	Yes	Yes	Yes	ND
Rode	01	30	PC	Complete	Yes	Yes	Yes	Yes	ND
Kim	01	31	RC	Complete	ND	Yes	Yes	Yes	ND
Mortelè	01	32	RC	Complete	No	No	Yes	Yes	ND
Trevisani	01	33	CC	Complete	ND	No	Yes	Yes	ND
Krinsky	01	34	PC	Complete	ND	ND	Yes	Yes	ND
Bennett	02	35	RC	Complete	ND	ND	Yes	Yes	ND
de Ledinghen	02	36	PC	Complete	Yes	ND	Yes	Yes	ND
Zacherl	02	37	PC	Complete	Yes	Yes	Yes	Yes	ND
Barthia	02	38	PC	Complete	Yes	ND	Yes	Yes	ND
Mori	01	39	PC	Complete	Yes	Yes	Yes	Yes	Yes
Teefey	03	40	PC	Complete	Yes	ND	Yes	Yes	Yes
Tong	01	41	PC	Partial	ND	Yes	Yes	Yes	ND
Libbrecht	02	42	PC	Complete	Yes	Yes	Yes	Yes	ND
Burrel	03	43	PC	Complete	Yes	Yes	Yes	Yes	ND
Nguyen	02	44	CC	Partial	ND	Yes	No	Yes	ND
Marrero	03	45	CC	Partial	ND	Yes	Yes	Yes	ND

PC = prospective cohort; CC = case-control; RC = retrospective cohort; ND = not defined.

to improve the accuracy of a diagnosis of HCC, did not fulfill our inclusion criteria. Having discussed the possible sources of heterogeneity, we shall now consider the pooled estimate of US operative characteristics, which revealed poor sensitivity (60%, 95% CI 44–76) and unexpectedly high specificity (97%, 95% CI 95–98). Interestingly, although their specificity was similar (97%), the eight studies using the pathology of the explanted liver as the reference standard (21, 23, 27, 30, 31, 35, 40, 42) were even less sensitive (48%). These can be confidently considered of high quality, as they consecutively enrolled patients with end-stage liver disease and the reference standard was always the same, thus avoiding the work-up bias (Table 4). The low sensitivity level of this subset of studies may be related to their almost perfect reference standard, because the pathological examination of an explanted liver allowed to detect even sub-centimetric neoplastic nodules in an advanced nodular pattern. Furthermore, the sensitivity of US in detecting focal lesions may be further decreased in patients with end-stage liver disease (*i.e.*, those undergoing OLT) and a severely shrunken liver parenchyma. On the other hand, the unexpectedly high specificity of the studies as a whole and those in the above subset could be attributable to the concomitant increased incidence of HCC and decreased prevalence of benign focal lesions (mainly hemangiomas) in patients with cirrhosis (49), as has recently

been confirmed by data indicating that at least 50% of the hemangioma-like lesions detected by US in patients with advanced cirrhosis are actually HCCs (50). Despite its high degree of specificity (>95%), the present data do not support the use of US as a confirmative diagnostic tool in current medical practice. In a clinical setting of patients with chronic liver disease, we have estimated a pre-test probability of HCC of about 10%, somewhere between the 7% obtained at baseline in a large prospective surveillance study including Child class A cirrhotics (51) and the 17% observed in a large series of patients undergoing liver transplantation (52). The pooled LR+ value of 18, increased the post-test probability of HCC to 66%, but this is still too low to support the confirmatory use of US, especially when a surgical procedure is scheduled; furthermore, the pooled LR– value of 0.52 does not support an exclusion strategy because the post-test probability only decreases from 10% to 5%.

The operative characteristics of spiral CT were comparable with those of US (see Tables 3 and 4). Although all of the studies used the pathology of the explanted liver as the reference standard, the variability of their results indicate the presence of a heterogeneity that is probably related to differences in technical details (*e.g.*, variations in collimation and the contrast media infusion rate) and image interpretation, although no pertinent data were available.

The pooled MRI sensitivity (81%) and specificity (85%) estimates were, respectively, higher and lower than those obtained with US or CT and, once again, the wide range of the results obtained in OLT candidates may be attributable to differences in equipment and technical details. Surprisingly, one study mainly designed to assess the role of MRI angiography in HCC staging (43) showed absolute sensitivity (100%) and a high degree of specificity (95%), possibly because of patient selection; 29 of the 50 patients had a previously diagnosed HCC, that represented the indication for OLT.

Interestingly, although AFP is widely used in screening or surveillance programs and clinical practice, only nine studies fulfilled our inclusion criteria (20, 22, 25–27, 33, 41, 44, 45). As expected, the AFP operative characteristics varied depending on the cutoff value (see Fig. 4), but the heterogeneity of the results cannot be entirely attributed to the threshold effect because, as shown in Figure 5, the sensitivity and specificity estimates for the same cutoff value were clearly different. In addition, five of the studies involved case-control series, thus leading to a possible overestimate of diagnostic accuracy (6), and only one (27) assessed the accuracy of AFP against the pathology of the explanted liver. In a recent systematic review, focused on the accuracy of AFP in detecting HCC in patients with chronic hepatitis C, concerns about the quality of the studies were confirmed, as no summary estimates of sensitivity and specificity were obtainable (53).

The results of our review could have a considerable impact on current clinical practice, in which all of these tests are usually employed. The suggestion of using hepatic CT and/or MRI to confirm US findings consistent with a possible HCC (5), can be challenged because these tests (as imaging techniques) cannot be regarded as *a priori* conditionally independent (which prevents any easy prediction of the diagnostic yield of their sequential use). In fact, no increase in diagnostic accuracy was observed when US and CT findings were combined in the only two studies which assessed this critical issue (23, 27). However, contrast enhanced techniques (US, CT, and MRI) could detect some characteristic patterns (*e.g.*, an arterial enhancement followed by wash out) which could reflect neoplastic vascularization, and thus improve their diagnostic yield as suggested by preliminary data (54).

In relation to the combination of two independent tests (*e.g.*, AFP levels and one imaging technique), the use of an AFP cutoff level of >20 ng/mL improves the sensitivity but not the specificity of US and CT (27); the only way of improving specificity to a degree that would be useful for a confirmatory strategy is to increase the cutoff value, but this greatly reduces sensitivity (Fig. 2).

Recent data, from different studies focused on patients receiving OLT, showed a low preoperative HCC detection; in fact, not only, a high rate of false negative results was found, as expected, but also a surprisingly high false positive rate ($>30\%$) (55, 56). Furthermore, in another recent study using histologic examination of hypovascular liver lesions (*i.e.*, those without post-contrast enhancement), HCC was found in

up to 38% of the cases (57), thus confirming the poor accuracy of diagnostic techniques in HCC detection, clearly evidenced by this review. Therefore, a definite diagnosis of HCC still relies in many cases on the use of US- or CT-guided liver biopsy and, in addition, there is clear evidence that any US-detected focal lesion in patients with liver cirrhosis should be regarded as a potential HCC and investigated further (50).

It is worth noting that none of the above tests seems to be sufficiently sensitive to support an effective surveillance program for the early detection of HCC in cirrhosis. The conflicting results of such programs (58–60) may be at least partially due to the inability of diagnostic tests to detect more than 60% of neoplasms, and not only to the lack of effective treatment. It is interesting that the pooled estimate of MRI sensitivity is more than 80% and has increased over the years, because this supports its possible future role in surveillance programs.

This review inevitably has a number of limitations. According to Feinstein (61), a diagnostic test should not only confirm the presence or absence of a given disease, but may also be useful in staging it or detecting the related risk factors and/or concomitant diseases. We did not consider the usefulness of noninvasive techniques in staging HCC. Furthermore, the potential accuracy of the imaging techniques (US, CT, and MRI) and AFP concentrations may have been negatively affected by our decision to consider explanted livers as one of the reference standards; in addition, it should also be pointed out that the clinical meaning of sub-centimetric HCC nodules in explanted livers has still not been defined. Another critical point may be that the lack of sufficient data prevented a correct subgroup analysis that may have identified the possible sources of the heterogeneous results. Moreover, we were unable to assess the reproducibility of the diagnostic techniques, of relevance especially in the case of US, as most of the included studies did not actually assess this critical issue (see Table 4 for details).

One further limitation is the possibility of publication bias because, although methods for dealing with it have been developed, their applicability to the assessment of diagnostic tests has not yet been defined and it is therefore clearly possible that our pooled estimates are too optimistic insofar, as studies with favorable results are more likely to be submitted and published. Nevertheless, as from Figure 6, our funnel plot analysis revealed a possible publication bias only in the case of the AFP studies.

Finally, in this review we were unable to assess the cost-effective issue, even if it is well-known that MRI and CT currently are the most expensive techniques. Even the potential adverse effects of the diagnostic techniques (*e.g.*, radiation exposure at CT) were not assessed.

The results of this systematic review support the use of US as a specific diagnostic tool at least in patients with underlying cirrhosis and a high probability of HCC, but its poor sensitivity probably means that it is inadequate for screening purposes. The role of CT and MRI remains to be defined, particularly as additional tests aimed at confirming positive US

results but, if confirmed, the greater sensitivity of MRI may support its use as a screening test. Finally, the small number of AFP studies, together with their heterogeneous estimates and poor quality, clearly indicate the need for high-quality prospective studies.

ACKNOWLEDGMENTS

The authors would like to thank the *Associazione Amici della Gastroenterologia del Granelli* (AAGG) for its continuing support, and the CARIPLO Foundation for a special grant.

Reprint requests and correspondence: Dario Conte, M.D., Postgraduate School of Gastroenterology, Padiglione Granelli 3° piano, Fondazione IRCCS—Ospedale Maggiore, Mangiagalli e Regina Elena, Via F. Sforza 35, 20122 Milano, Italy.

Received August 1, 2005; accepted November 7, 2005.

REFERENCES

- National Institutes of Health. National Institutes of Health Consensus Development Conference Statement: Management of hepatitis C: 2002—June 10–12, 2002. *Hepatology* 2002;36:S3–20.
- Lauer M, Walker BD. Hepatitis C virus infection. *N Engl J Med* 2001;345:41–52.
- Ishak K, Baptista A, Bianchi L, et al. Histological grading and staging of chronic hepatitis. *J Hepatol* 1995;22:696–9.
- Llovet JM, Fuster J, Bruix G, for the Barcelona Clinic Liver Cancer (BCLC) group. Intention-to-treat of surgical treatment for early hepatocellular carcinoma: Reception *versus* transplantation. *Hepatology* 1999;30:1434–40.
- Bruix J, Sherman M, Llovet JM, et al. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. *J Hepatol* 2001;35:421–30.
- Sackett D, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, ed. *The evidence base of clinical diagnosis*. London: BMJ Publishing Group, 2002:19–38.
- Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;274:645–51.
- Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061–6.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
- Deville WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Med Res Methodol* 2002;2:9–22.
- Agresti A. *An introduction to categorical data analysis*. New York: John Wiley & Sons, 1997:39–45.
- Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials* 1986;7:267–75.
- Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic summary point estimates. *Med Decis Making* 1993;13:253–7.
- 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;12:1293–316.
- Begg C. Publication bias. In: Xoooper H, Hedges L, eds. *The handbook of research synthesis*. New York: Russel Sage Foundation, 1994:399–409.
- Okazaki N, Yoshida T, Yoshino M, et al. Screening of patients with chronic liver disease for hepatocellular carcinoma by ultrasonography. *Clin Oncol* 1984;10:241–6.
- Maringhini A, Cottone M, Sciarrino E, et al. Ultrasonographic and radionuclide detection of hepatocellular carcinoma in cirrhotics with low alpha-fetoprotein levels. *Cancer* 1984;15(54):2924–6.
- Kobayashi K, Sugimoto T, Makino H, et al. Screening methods for early detection of hepatocellular carcinoma. *Hepatology* 1985;5:1100–5.
- Tanaka S, Kitamura T, Ohshima A, et al. Diagnostic accuracy of ultrasonography for hepatocellular carcinoma. *Cancer* 1986;58:344.
- Piantino P, Arrigoni A, Brunetto MR, et al. Alpha-fetoprotein in hepatic pathology and hepatocarcinoma. *J Nucl Med Allied Sci* 1989;33:34–8.
- Dodd GD, Miller WJ, Baron RL, et al. Detection of malignant tumors in end-stage cirrhotic livers: Efficacy of sonography as a screening technique. *AJR Am J Roentgenol* 1992;159:727–33.
- Lopez JB, Balasegaram M, Thambyrajah V. Serum CA 125 as a marker of hepatocellular carcinoma. *Int J Biol Markers* 1996;11:178–82.
- Saada J, Bhattacharya S, Dhillon AP, et al. Detection of small hepatocellular carcinomas in cirrhotic livers using iodised oil computed tomography. *Gut* 1997;41:404–7.
- Born M, Layer G, Kreft B, et al. MRT, CT und in der diagnostik maligner lebertumoren bei leberzirrhose. *Fortschr Rontgenstr* 1998;567–72.
- Bayati N, Silverman AL, Gordon SC. Serum alpha-fetoprotein levels and liver histology in patients with chronic hepatitis C. *Am J Gastroenterol* 1998;93:2452–6.
- Chalasan N, Horlander JC Sr., Said A, et al. Screening for hepatocellular carcinoma in patients with advanced cirrhosis. *Am J Gastroenterol* 1999;94:2988–93.
- Gambarin-Gelwan M, Wolf DC, Shapiro R, et al. Sensitivity of commonly available screening tests in detecting hepatocellular carcinoma in cirrhotic patients undergoing liver transplantation. *Am J Gastroenterol* 2000;95:1535–8.
- Lim JH, Kim CK, Lee WJ, et al. Detection of hepatocellular carcinomas and dysplastic nodules in cirrhotic livers: Accuracy of helical CT in transplant patients. *AJR Am J Roentgenol* 2000;175:693–8.
- Peterson MS, Baron RL, Marsh JW Jr, et al. Pretransplantation surveillance for possible hepatocellular carcinoma in patients with cirrhosis: Epidemiology and CT-based tumor detection rate in 430 cases with surgical pathologic correlation. *Radiology* 2000;217:743–9.
- Rode A, Bancel B, Douek P, et al. Small nodule detection in cirrhotic livers: Evaluation with US, spiral CT, and MRI and correlation with pathologic examination of explanted liver. *J Comput Assist Tomogr* 2001;25:327–36.
- Kim CK, Lim JH, Lee WJ. Detection of hepatocellular carcinoma and dysplastic nodules in cirrhotic liver. *J Ultrasound Med* 2001;99–104.
- Mortele KJ, de Keuleleire K, Praet M, et al. Malignant focal hepatic lesions complicating underlying liver disease: Dual-phase contrast-enhanced spiral CT sensitivity and specificity in orthotopic liver transplant patients. *Eur Radiol* 2001;11:1631–8.
- Trevisani F, D'Intino PE, Morselli-Labate AM, et al. Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: Influence of HBsAg and anti-HCV status. *J Hepatol* 2001;34:570–5.

34. Krinsky GA, Lee VS, Theise ND, et al. Hepatocellular carcinoma and dysplastic nodules in patients with cirrhosis: Prospective diagnosis with MR imaging and explantation correlation. *Radiology* 2001;219:445–54.
35. Bennett GL, Krisky GA, Abitbol RJ, et al. Sonographic detection of hepatocellular carcinoma and dysplastic nodules in cirrhosis: Correlation of pretransplantation sonography and liver explant pathology in 200 patients. *AJR Am J Roentgenol* 2002;179:75–80.
36. de Ledinghen V, Laharie D, Lecesne R, et al. Detection of nodules in liver cirrhosis: Spiral computed tomography or magnetic resonance imaging? A prospective study of 88 nodules in 34 patients. *Eur J Gastroenterol Hepatol* 2002;14:159–65.
37. Zacherl J, Pokieser P, Wrba F, et al. Accuracy of multiphase helical computed tomography and intraoperative sonography in patients undergoing orthotopic liver transplantation for hepatoma: What is the truth? *Ann Surg* 2002;235:528–32.
38. Bhartia B, Ward J, Guthrie JA, et al. Hepatocellular carcinoma in cirrhotic livers: Double-contrast thin-section MR imaging with pathologic correlation of explanted tissue. *AJR Am J Roentgenol* 2003;180:577–84.
39. Mori K, Scheidler J, Helmberger T, et al. Detection of malignant hepatic lesions before orthotopic liver transplantation: Accuracy of ferumoxides-enhanced MR imaging. *AJR Am J Roentgenol* 2002;179:1045–51.
40. Teefey SA, Hildeboldt CC, Dehdashti F, et al. Detection of primary hepatic malignancy in liver transplant candidates: Prospective comparison of CT, MR imaging, US, and PET. *Radiology* 2003;226:533–42.
41. Tong MJ, Blatt LM, Kao VW. Surveillance for hepatocellular carcinoma in patients with chronic viral hepatitis in the United States of America. *J Gastroenterol Hepatol* 2001;16:553–9.
42. Libbrecht L, Bielen D, Verslype C, et al. Focal lesions in cirrhotic explant livers: Pathological evaluation and accuracy of pretransplantation imaging examinations. *Liver Transpl* 2002;8:749–61.
43. Burrell M, Llovet JM, Ayuso C, et al. MRI angiography is superior to helical CT for detection of HCC prior to liver transplantation: An explant correlation. *Hepatology* 2003;38:1034–42.
44. Nguyen MH, Garcia RT, Simpson PW, et al. Racial differences in effectiveness of alpha-fetoprotein for diagnosis of hepatocellular carcinoma in hepatitis C virus cirrhosis. *Hepatology* 2002;36:410–7.
45. Marrero JA, Su GL, Wei W, et al. Des-gamma carboxyprothrombin can differentiate hepatocellular carcinoma from nonmalignant chronic liver disease in American patients. *Hepatology* 2003;37:1114–21.
46. Mok TS, Yu SC, Lee C, et al. False-negative rate of abdominal sonography for detecting hepatocellular carcinoma in patients with hepatitis B and elevated serum α -fetoprotein levels. *AJR Am J Roentgenol* 2004;183:453–8.
47. Lencioni R, Pinto F, Armillotta N, et al. Assessment of tumor vascularity in hepatocellular carcinoma: Comparison of power Doppler US and color Doppler US. *Radiology* 1996;201:353–8.
48. Lencioni R, Cioni D, Bartolozzi C. Tissue harmonic and contrast-specific imaging: Back to gray scale in ultrasound. *Eur Radiol* 2002;12:151–65.
49. Brancatelli G, Federle MP, Blachar A, et al. Hemangioma in the cirrhotic liver: Diagnosis and natural history. *Radiology* 2001;219:69–74.
50. Caturelli E, Bartolucci F, Biasini E, et al. Diagnosis of liver nodules observed in chronic liver disease patients during ultrasound screening for early detection of hepatocellular carcinoma. *Am J Gastroenterol* 2002;97:397–405.
51. Colombo M. Hepatocellular carcinoma. In: McDonald J, Burroughs A, Feagan B, eds. Evidence-based gastroenterology and hepatology. London: BMJ Books, 2004:517–25.
52. Mion F, Grozel L, Boillot O, et al. Adult cirrhotic liver explants: Precancerous lesions and undetected small hepatocellular carcinomas. *Gastroenterology* 1996;111:1587–92.
53. Gupta S, Bent S, Kohlwes J. Test characteristics of alpha-fetoprotein for detecting hepatocellular carcinoma in patients with hepatitis C. A systematic review and critical analysis. *Ann Intern Med* 2003;139:46–50.
54. Marrero JA, Hussain HK, Nghiem HV, et al. Improving the prediction of hepatocellular carcinoma in cirrhotic patients with an arterially-enhancing liver mass. *Liver Transpl* 2005;281–9.
55. Wiesner RH, Freeman RB, Mulligan DC. Liver transplantation for hepatocellular cancer: The impact of the MELD allocation policy. *Gastroenterology* 2004;127:S261–7.
56. Hayashi PH, Trotter JF, Forman L, et al. Impact of pretransplant diagnosis of hepatocellular carcinoma on cadaveric liver allocation in the era of MELD. *Liver Transpl* 2004;10:42–8.
57. Bolondi L, Gaiani S, Celli N, et al. Characterization of small nodules in cirrhosis by assessment of vascularity: The problem of hypovascular hepatocellular carcinoma. *Hepatology* 2005;42:27–34.
58. Colombo M, de Franchis R, Del Ninno E, et al. Hepatocellular carcinoma in Italian patients with cirrhosis. *N Engl J Med* 1991;325:675–80.
59. Sangiovanni A, Del Ninno E, Fasani P, et al. Increased survival of cirrhotic patients with a hepatocellular carcinoma detected during surveillance. *Gastroenterology* 2004;126:1005–14.
60. Wun YT, Dickinson JA. Alpha-fetoprotein and/or liver ultrasonography for liver cancer screening in patients with chronic hepatitis B. *Cochrane Database Syst Rev* 2003;(2):CD002799.
61. Feinstein J. Misguided efforts and future challenges for research on “diagnostic tests.” *Epidemiol Community Health* 2002;56:330–2.