The version of QUADAS-2 used in this assessment splits domain 2 into 'index test' and 'comparator' and includes additional signalling questions to accommodate primary studies that assess multiple tests. Only the 'patient selection' domain includes an applicability domain as it was considered that the inclusion criteria matched the review question for the 'index test', 'comparator' and 'reference standard' domains.

Before starting the risk of bias assessment we considered the relevance of each signalling question to our review, as well as the potential need for additional questions. Further criteria were then defined, as needed, to ensure consistent application of signalling questions and to help in the judgement of the risk of bias. Many signalling questions were not further specified and the answer was judged to be 'yes' if it was clearly reported in the study. If the answer to a signalling question was not clearly reported the question was judged as 'unclear' unless specified differently. 'No' was answered if it was clear from the reporting that an aspect was not fulfilled. Details of the assessment criteria used are reported below.

Domain 1: patient selection

Risk of bias

Question 1: Was a consecutive or random sample of patients enrolled?

- 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' = high risk of bias.

Question 2: Was a case–control design avoided?

- 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' = high risk of bias.

Question 3: Did the study avoid inappropriate exclusions?

- 'no' for < 10% of patients or 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' for ≥ 10% of patients = high risk of bias.

Concerns regarding applicability

- Included patients were adults with FLLs with uncertain diagnosis on standard US or other imaging modalities = 'low concern'.
- Included patients were adults with known liver malignancy who were being assessed for recurrence or response to treatment = 'low concern'.
- Included patients were adults with FLLs detected on standard US or other imaging, in which it was not clear if these examinations were diagnostic = 'unclear concern'.

Domain 2a: index test

Risk of bias

Question 1: Were the index test results interpreted without knowledge of the results of the reference standard?
Question 2: Were the index test results interpreted without knowledge of the comparator?
Question 3: Did the study prespecify the threshold for a positive result?

The same criteria applied to each of the three signalling questions:

- 'yes' = low risk of bias
- 'unclear'' = unclear risk of bias
- 'no' = high risk of bias.

Domain 2b: comparator test

Risk of bias

Question 1: Were the comparator test results interpreted without knowledge of the results of the reference standard?
Question 2: Were the comparator test results interpreted without knowledge of the index test?
Question 3: Did the study prespecify the threshold for a positive result?

The same criteria applied to each of the three signalling questions:

- 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' = high risk of bias.

Domain 3: reference standard

Risk of bias

Question 1: Is the reference standard likely to correctly classify the target condition?

- 'yes' if ≥90% of test results were confirmed using the reference standard specified by the inclusion criteria (pathology for test positive and pathology or minimum 6 months' follow-up for test negative) = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' if <90% of test results were confirmed using the reference standard specified by the inclusion criteria (pathology for test positive and pathology or minimum 6 months' follow-up for test negative) = high risk of bias.

Question 2: Were the reference standard results interpreted without knowledge of the results of the index test?
Question 3: Were the reference standard results interpreted without knowledge of the results of the comparator test?

The same criteria applied to signalling questions 2 and 3:

- 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' = high risk of bias.

Domain 4: flow and timing

Question 1: Was there an appropriate interval between index test and reference standard?

The time interval between index and reference standard (pathology) had to be ≤1 month to be judged as 'adequate' and follow-up had to be ≥6 months to be judged as 'adequate'.

- 'no' for <10% of patients or 'yes' = low risk of bias
- the answer was judged to be 'unclear' if the time interval was not reported or if it was unclear what proportion of patients had an inadequate time interval between index test and reference standard = unclear risk of bias
- 'no' for ≥10% of patients = high risk of bias.

Question 2: Was there an appropriate interval between comparator test and reference standard?

The time interval between index and reference standard (pathology) had to be ≤1 month to be judged as 'adequate' and follow-up had to be ≥6 months to be judged as 'adequate'.

- 'no' for <10% of patients or 'yes' = low risk of bias
- the answer was judged to be 'unclear' if the time interval was not reported or if it was unclear what proportion of patients had an inadequate time interval between index test and reference standard = unclear risk of bias
- 'no' for ≥10% of patients = high risk of bias.

Question 3: Was there an appropriate interval between index test and comparator test?

The time interval between index and comparator had to be ≤1 month to be judged as 'adequate'.

- 'no' for <10% of patients or 'yes' = low risk of bias
- the answer was judged to be 'unclear' if the time interval was not reported or if it was unclear what proportion of patients had an inadequate time interval between index test and reference standard = unclear risk of bias
- 'no' for ≥10% of patients = high risk of bias.

Question 4: Did all patients receive a reference standard?

- 'no' for <10% of patients or 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' for ≥10% of patients = high risk of bias.

Question 5: Did all patients receive the same reference standard?

Acceptable reference standards were defined separately for test-positive and test-negative patients. The following criteria are therefore applied separately to test-positive and test-negative patients:

- 'no' for <10% of test-positive patients and <10% of test-negative patients, or 'yes' = low risk of bias
- 'unclear' = unclear risk of bias
- 'no' for ≥10% of test-positive or test-negative patients = high risk of bias.

Question 6: Were all patients included in the analysis?

- 'no' for <10% of patients or 'yes' = low risk of bias
- 'unclear' = unclear risk of bias

- ‘no’ for ≥ 10% of patients = high risk of bias.

The following criteria were used to reach a per-domain judgement of risk of bias:

- if at least one of the signalling questions of a domain had an answer associated with a high risk of bias the domain was judged to have a high risk of bias
- if the answer to any of the signalling questions was ‘unclear’ and the answers to the remaining questions were ‘yes’, the risk of bias was judged to be unclear
- the answer to all the signalling questions had to be ‘yes’ for the domain to be judged as having a low risk of bias.