Accuracy of Magnetic Resonance Imaging in Diagnosis of Liver Iron Overload: a Systematic Review and Meta-analysis

Short title: MRI for Liver Iron Overload

MARIA SARIGIANNI,*‡ ARIS LIAKOS,* EFTHYMIA VLACHAKI,* PASCHALIS PASCHOS,§ ELENI ATHANASIADOU,* VICTOR M. MONTORI,‡ MOHAMMAD HASSAN MURAD,‡ and APOSTOLOS TSAPAS*||

*Clinical Research and Evidence-Based Medicine Unit, Aristotle University Thessaloniki, Thessaloniki, Greece; ‡Knowledge and Evaluation Research Unit, College of Medicine, Mayo Clinic, Minnesota, USA; §Thalassemia Unit, Second Medical Department, Aristotle University Thessaloniki, Thessaloniki, Greece; and ||Harris Manchester College, University of Oxford, Oxford, United Kingdom

Funding: Supported in part by a post-graduate research grant from the Aristotle University Thessaloniki Research Committee, Greece (M.S.).

Abbreviations used in this paper: CI, confidence interval; GRE, gradient-recalled echo; LR, likelihood ratio; MRI, magnetic resonance imaging; SE, spin echo; SROC, summary receiver operating characteristic.

Correspondence: Apostolos Tsapas, MD, MSc, PhD, Clinical Research and Evidence-Based Medicine Unit, Second Medical Department, Aristotle University
Thessaloniki, Hippokratio General Hospital, 49 Konstantinoupolis Street, Thessaloniki, 54642, Greece. e-mail: atsapas@auth.gr; phone: (30) 2310992937; fax: (30) 2310992937.

Conflicts of interest: The authors declare no competing interests.

Writing assistance: No writing assistance was provided.

Author contributions: M.S., E.V., V.M., M.M. and A.T. were responsible for study concept and design. M.S., P.P., A.L. and E.A. performed the study selection, assessment of methodological quality, data extraction, and statistical analysis. M.S., P.P., A.L. and A.T. prepared the first draft of the report, which was critically revised by all authors. All authors approved the final version of the manuscript.
Abstract

BACKGROUND & AIMS: Guidelines advocate use of magnetic resonance imaging (MRI) to estimate concentrations of iron in liver, identify patients with iron overload, and guide titration of chelation therapy. However, this recommendation was not based on a systematic synthesis and analysis of the evidence for MRI’s diagnostic accuracy.

METHODS: We conducted a systematic review and meta-analysis to investigate the diagnostic accuracy of MRI in identifying liver iron overload in patients with hereditary hemochromatosis, hemoglobinopathy, or myelodysplastic syndrome; liver biopsy analysis was used as the reference standard. We searched MEDLINE and EMBASE databases, the Cochrane Library, and gray literature, and computed summary receiver operating curves by fitting hierarchical models. We assessed methodologic quality using the Quality Assessment of Diagnostic Accuracy Studies 2 tool.

RESULTS: Our final analysis included 20 studies (819 patients, total). Sensitivity and specificity values varied greatly, ranging from 0.00 to 1.00 and from 0.50 to 1.00, respectively. Due to substantial heterogeneity and variable positivity thresholds, we calculated only summary receiver operating curves (and summary estimate points for studies that used the same MRI sequences). T2 spin echo and T2* gradient-recalled echo MRI sequences accurately identified patients without liver iron overload (liver iron concentration > 7 mg Fe/g dry liver weight) (negative likelihood ratios, 0.10 and 0.05 respectively). However, these MRI sequences are less accurate in establishing a definite diagnosis of liver iron overload (positive likelihood ratio of 8.85 and 4.86 respectively).

CONCLUSIONS: Based on a meta-analysis, measurements of liver iron concentration by MRI may be accurate enough to rule out iron overload, but not to
definitely identify patients with this condition. Most studies did not use explicit and pre-specified MRI thresholds for iron overload, so some patients may have been inaccurately diagnosed with this condition. More studies are needed of standardized MRI protocols and to determine the effects of MRI surveillance on development of chronic liver disease and patient survival.

**KEY WORDS:** diagnostic accuracy meta-analysis, sickle cell disease, thalassemia, LIC
**Introduction**

Iron overload is a principal determinant of major complications in patients with hereditary hemochromatosis or conditions requiring chronic transfusions, including thalassemia, sickle cell disease and myelodysplastic syndrome.\(^1\) Assessment of total body iron stores is used to tailor titration of chelation therapy in patients with transfusional iron overload, or to guide the diagnostic approach of suspected hereditary hemochromatosis in C282Y negative cases.\(^2\) Liver iron concentration is a reliable marker of total body iron and the gold standard for estimating liver iron concentration is liver biopsy.\(^3\)-\(^7\) However, it is an invasive method potentially associated with significant complications.\(^8\)

In the past, several studies have assessed the diagnostic accuracy of Magnetic Resonance Imaging (MRI) for estimation of liver iron concentration in patients with hereditary hemochromatosis or transfusional iron overload.\(^6\), \(^9\)-\(^12\) Existing guidelines advocate use of MRI for assessment of total body iron in patients with thalassemia major or post-transfusional iron overload,\(^7\), \(^13\), \(^14\) nevertheless this recommendation is not based on a systematic synthesis and appraisal of the available evidence-base regarding its diagnostic accuracy.

We conducted a systematic review and meta-analysis to investigate the diagnostic accuracy of MRI for estimating liver iron overload in patients with hereditary hemochromatosis, hemoglobinopathies (sickle cell disease and thalassemia) and myelodysplastic syndrome. Our secondary objective was to explore the effect of various MRI characteristics on its diagnostic accuracy.

**Materials and Methods**
We report this systematic review according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses statement\textsuperscript{15}.

**Eligibility Criteria**

We considered all studies that assessed the diagnostic accuracy of MRI for liver iron concentration in patients with hereditary hemochromatosis, hemoglobinopathies (sickle cell disease and thalassemia) or myelodysplastic syndrome. We included all cohort and case-control studies that used $\geq 1.5$ Tesla scanner systems and utilized liver biopsy with chemical quantification of liver iron concentration as reference standard.\textsuperscript{3} Lower field strength scanners cannot accurately detect iron deposition in the liver and have ceased being used in the clinical setting.\textsuperscript{16,17} We excluded studies in which we could not reconstruct a $2 \times 2$ diagnostic table.

**Search Strategy**

We designed a comprehensive search strategy of several electronic databases and gray literature sources, without imposing any date or language restrictions. We searched MEDLINE via PubMed, EMBASE via Ovid, Web of Science and Cochrane Library utilizing both free-text terms and relevant Medical Subject Headings or Emtree terms. We also hand-searched trial registries and abstracts from relevant scientific meetings. Finally, we scanned Web sites of companies manufacturing pertinent MRI-based technologies and perused reference lists of included studies. A detailed search strategy and list of literature sources searched are found in the Supplementary Material. Last search was run on June 14, 2013.

**Study Selection and Data Extraction**

Following deduplication, two authors working independently reviewed all titles and abstracts. Selected articles were screened for eligibility in full text independently by the same reviewers. Disagreement at any stage was arbitrated by a third reviewer.
Using a predesigned form, three authors extracted in duplicate data from eligible studies. In case of studies where only a subgroup of participants met the review inclusion criteria, data were extracted and presented only for that particular subgroup. For studies with multiple reports we collated data to maximize yield of information. When several reports for the same study population were available (duplicate or cumulative publication), we used the most up-to-date or complete report, to avoid double counting of participants. Finally, we contacted authors of original studies for missing or unclear data.

We extracted data to reconstruct 2 × 2 tables for each of the clinically relevant liver iron concentration values (> 2, > 7 or > 15 mg Fe/g dry liver weight) and the corresponding MRI positivity thresholds for every individual study. Liver iron concentration > 2 mg Fe/g dry liver weight is used to establish the diagnosis of iron overload, > 7 mg Fe/g dry liver weight suggests increased risk for iron-induced complications and is the threshold utilized to initiate or intensify chelation therapy, and finally > 15 mg Fe/g dry liver weight is associated with substantial risk for cardiac disease and early death.\(^{18}\) If a 2 × 2 table was not available and studies reported results with correlation plots, we extracted raw data utilizing Plot digitizer (http://plotdigitizer.sourceforge.net). For studies that did not provide a 2 × 2 table and raw data were available, we calculated MRI positivity thresholds for every individual biopsy threshold (2, 7 or 15 mg Fe/dry weight of liver) based on a study-specific linear regression line.

Assessment of Methodological Quality

We used the Quality Assessment of Diagnostic Accuracy Studies 2 tool to assess the quality of included studies.\(^ {19}\) Two review authors formulated the assessment questions, considered the signaling questions for relevance to the review
question and agreed on a review-specific version of the tool. Risk of bias and applicability were assessed independently in duplicate for all included reports by two reviewers, and disagreements were resolved by consensus. Assessment for every domain was based on an a priori formulated rule that the overall risk of bias was equal with the highest risk of bias for any of the respective signaling questions.

**Statistical Analysis and Data Synthesis**

We used $2 \times 2$ tables to calculate sensitivity and specificity for every individual study utilizing RevMan 5.2 (Nordic Cochrane Center, Copenhagen, Denmark). We fitted hierarchical models to depict summary receiver operating characteristic (SROC) curves using user-written modules (metandi and midas) in Stata 12.1 (Stata Corporation, College Station, Texas).\textsuperscript{20,21} We applied a continuity correction of 0.5 in studies with zero cells. When appropriate, we identified the average operating point and computed average sensitivities, specificities, likelihood ratios of positive (LR+) and negative test results (LR–). We assessed statistical heterogeneity using the $I^2$ and Chi$^2$ statistics, with $I^2$ values higher than 50% or $P < .10$ respectively representing high heterogeneity.\textsuperscript{22}

For the main analysis we utilized data from all studies reporting diagnostic accuracy for the > 7 mg Fe/g dry liver weight biopsy threshold, regardless of MRI sequence. We also conducted sensitivity analyses for the remaining clinically relevant biopsy thresholds (> 2 and > 15 mg Fe/g dry liver weight). When results for multiple MRI sequences were available within the same study, we utilized data for the most common MRI sequence across studies. To explore heterogeneity, we conducted subgroup analyses based on MRI sequence. We undertook a meta-regression analysis on specificity and sensitivity in the bivariate model, using as predictors publication year and prevalence of iron overload (liver iron concentration > 7 mg Fe/g dry liver
weight) among included patients with hereditary hemochromatosis, hemoglobinopathies or myelodysplastic syndrome. We planned to perform a sensitivity analysis based on methodological quality, excluding studies at high risk of bias. We depicted conditional probability plots for practicing physicians to interpret the clinical utility of MRI sequences.

Finally, we assessed risk of bias across studies, exploring for small study effects (publication bias). We depicted a funnel plot and tested for asymmetry by means of linear regression analysis of diagnostic log odds ratio against effective sample size, with value $P < .10$ suggesting significant asymmetry.  

**Results**

**Results of Search and Characteristics of Included Studies**

Our search retrieved a total of 7319 records from all literature sources. Flow diagram of study selection process is depicted in Figure 1. Following deduplication, we screened 4837 reports in title and abstract form and rejected 4744 reports as non-eligible. Inter-observer agreement was very good (kappa = 0.87). We assessed the remaining 93 reports in full-text and identified 20 eligible studies that were included in our systematic review. Authors of 11 studies responded to our request for additional data. For nine studies where only a subgroup of participants met the review inclusion criteria, we extracted data only for that particular subgroup.

The total number of patients assessed was 819. Number of eligible patients analyzed per study ranged from eight to 233 (median 26). Reported age ranged from three to 79 years (mean 31 years), while proportion of males ranged from 40% to
77%. All included studies utilized 1.5 Tesla MRI scanners. Characteristics of included studies are summarized in Table 1.

Methodological Quality of Included Studies

All included studies were considered at high risk of bias (Figure 2). Therefore, a planned sensitivity analysis based on methodological quality was not undertaken. Method of patient recruitment was unclear in eight studies (226 patients)11, 25, 27, 30, 33, 34, 36, 37 while the remaining 12 studies used a consecutive or random sample (594 patients).9, 10, 12, 16, 17, 24, 26, 28, 29, 31, 32, 35 MRI assessors were blinded to patients’ clinical details and biopsy results in 12 studies,9, 10, 12, 16, 17, 25, 29-33, 35 while the rest did not provide relevant data. The majority of included studies (n = 12) did not use an MRI positivity threshold and reported raw data or depicted results with correlation plots.9, 11, 24, 26-30, 32, 34, 36, 37 Five studies10, 12, 16, 17, 31 used an MRI positivity threshold which was not pre-specified, but was calculated based on data derived from the study population. Finally, two studies25, 33 did not clearly report whether the threshold used was pre-specified or not. Only one study35 used a pre-specified formula for interpretation of the index test results. All included studies used an acceptable reference standard. Reference standard results were interpreted without knowledge of index test results in nine studies,9, 10, 12, 16, 17, 24, 25, 29, 31 while in one study32 the pathologists were not blinded to index test results. The remaining studies did not report any information about blinding in the interpretation of biopsy results. Time interval between the index and reference test was less than one month in six studies,9, 12, 25, 29-31 between one and two months in four studies,11, 16, 32, 35 between three and six months in four studies,10, 24, 28, 33 between seven months and one year in two studies,17, 36 whereas four studies26, 27, 34, 37 did not provide relevant information. In eight studies (232 patients)9, 11, 12, 17, 24, 26, 30, 37 the reference standard was not applied in all patients,
hence these studies were rated at high risk for partial verification bias, whereas one study\textsuperscript{36} used two different reference standards. Finally, nine studies (466 patients)\textsuperscript{9, 11, 12, 24, 26, 29, 30, 35, 37} did not include all patients in the analysis.

For three studies\textsuperscript{17, 30, 31} there were concerns that the included patients do not match the review question, and for two studies\textsuperscript{16, 31} there were concerns that the index test or its interpretation differ from the review question. Finally, in all included studies there were no concerns regarding applicability based on the reference standard utilized.

**Diagnostic Accuracy**

In the main analysis (n = 17 studies), MRI sensitivity ranged from 0.00 to 1.00 (median 0.94). Specificity ranged from 0.50 to 1.00 (median 0.89) (Figure 3). We observed substantial heterogeneity between included studies, which could be attributed to the use of different MRI sequences and variable positivity thresholds. Hence, we decided to provide solely an SROC curve, and calculate summary estimate points only in subgroup analyses of studies that utilize the same MRI sequence\textsuperscript{38, 39}.

Pairs of observed values of sensitivity and specificity for MRI are presented in receiver operating characteristic space (Figure 4). The diagnostic odds ratio was 56 (95% CI 25, 127).

We sought to perform subgroup analyses for studies utilizing the same MRI sequence to identify patients requiring initiation or intensification of chelation therapy (> 7 mg Fe/g dry liver weight). We could pool results only for T2 spin echo (SE) and T2* gradient-recalled echo (GRE) or their reciprocals (R2 and R2* respectively) (Supplementary Figure 1). Diagnostic accuracy summary estimates for the aforementioned MRI sequences are shown in Supplementary Table 1. Both sequences had good diagnostic accuracy for detecting liver iron overload. We
depicted conditional probability plots to support decision making and interpretation of the clinical utility of these MRI sequences for detecting patients with liver iron concentration > 7 mg Fe/g dry liver weight (Figure 5). For example, in an average-risk population of patients with homozygous β-thalassemia and a pre-test probability (prevalence) of iron overload requiring chelation treatment (> 7 mg Fe/g dry liver weight) equal to 35%, MRI T2 SE increases the probability of iron overload to 83% when the test results are positive, and decreases the probability to 5% when the test results are negative. Similarly, MRI T2* GRE increases the probability of iron overload to 72% when the test results are positive, and decreases the probability to 3% when the test results are negative.

We also investigated the diagnostic accuracy of MRI across different reference standard thresholds (2 and 15 mg Fe/g dry liver weight). Results for the aforementioned biopsy cut-off values were highly inconsistent and were solely depicted using SROC curves (Supplementary Figure 2).

We conducted a meta-regression analysis to explore the potential effect of iron overload prevalence and publication year on heterogeneity. Using a threshold of $P < .10$ for statistical significance, prevalence of iron overload was a significant predictor of sensitivity ($P = .07$), however, it was not a significant predictor of specificity ($P = .43$). Publication year did not affect sensitivity or specificity estimates of MRI ($P = .91$ and .87 respectively).

We observed small study bias (possible publication bias) in the regression test for funnel plot asymmetry; $P$ value for the slope coefficient was .07 (Supplementary Figure 3).

Discussion
Liver biopsy is regarded as the reference standard for assessing LIC, nevertheless it is an invasive test potentially associated with risk for complications.\(^8\) Hence, existing guidelines support use of MRI for monitoring liver iron overload and titration of chelation therapy.\(^6,\,13\) We identified 20 studies utilizing many different MRI sequences with highly variable diagnostic accuracy. Sensitivity ranged from 0.00 to 1.00 (median 0.94) and specificity ranged from 0.50 to 1.00 (median 0.92). For our main analysis we did not pool results to calculate an overall summary estimate due to substantial heterogeneity. In subgroup analyses, both T2 SE and T2* GRE MRI sequences had good diagnostic accuracy (sensitivity 0.90 and 0.96, and specificity 0.87 and 0.80 respectively) for identifying patients at risk for iron-induced complications or requiring titration of chelation therapy (> 7 mg Fe/g dry liver weight). Hence, both methods can be used to safely rule out iron overload (LR− 0.10 and 0.05 respectively), nevertheless are moderately useful to establish a definite diagnosis (LR+ 8.85 and 4.86 respectively).\(^41\) Results for diagnostic accuracy at different iron concentration cut-off values (> 2 or > 15 mg Fe/g dry liver weight) were also highly inconsistent and could not be pooled reliably either.

Validity of our findings is seriously undermined by low quality of included studies, both at conduct and at reporting level. Most studies did not use the reference standard in all patients, hence are prone to verification bias. Moreover, the vast majority did not report 2 × 2 tables about the diagnostic accuracy of MRI for iron overload, but rather presented correlation plots without explicit and pre-specified index test (MRI) positivity thresholds. Furthermore, many studies did not provide adequate details regarding method of patient selection and blinding of MRI assessors. Future diagnostic trials should report results according to STAndards for Reporting of Diagnostic accuracy studies.\(^42\) Finally, we identified small study bias that could
indicate presence of publication bias. Nevertheless, the observed funnel plot asymmetry may as well be attributed to a variety of reasons other than publication bias, such as type of population studied and poor study quality, which could also interact with sample size and diagnostic accuracy estimates.\textsuperscript{23}

We should also acknowledge limitations at review level. For our review question, we decided to use liver biopsy as reference standard. Inhomogeneity of iron distribution within the liver in patients with cirrhosis may lead to sampling errors and variability in liver biopsy results,\textsuperscript{43} nevertheless liver biopsy remains the gold standard for assessment of liver iron concentration. Moreover, in order to tackle extensive use of correlation plots in the individual studies, we had to implement a data-driven approach to reproduce $2 \times 2$ tables. This may well lead to overestimation of diagnostic accuracy.\textsuperscript{44} Furthermore, we were not able to pool data in our main analysis, due to substantial heterogeneity. This could be attributed to high variability in MRI sequences or study populations and disease definitions over time.\textsuperscript{17, 27, 30, 32} We tried to explore heterogeneity utilizing subgroup and meta-regression analyses for the aforementioned limitations; we identified that prevalence was a significant predictor of sensitivity and specificity. We also sought to provide summary estimates across different biopsy cut-off values or MRI sequences utilized, but we were able to pool data only for T2 SE and T2* GRE sequences.

The strengths of this systematic review are related to the relevance of the clinical question addressed. We tried to explore the evidence-base of a diagnostic modality used in routine management of secondary iron overload, which is a common complication affecting a significant number of patients.\textsuperscript{6} Our review team comprised a variety of specialists, with expertise in the target condition and methodology of diagnostic accuracy systematic reviews. We used a comprehensive search strategy of
multiple electronic databases and gray literature sources to identify eligible studies. We utilized hierarchical models to synthesize available evidence. Finally, we explored validity of our analyses by means of subgroup and meta-regression analyses, and assessed the methodological quality of included studies using appropriate tools.19

Our review underlines the paucity of high quality studies designed to explore the diagnostic accuracy of MRI for liver iron overload. It is necessary for future trials to use a more rigorous design, have larger sample size, focus on specific sequences, and try to establish and utilize explicit MRI positivity thresholds. This will allow for reproducible results in clinical practice. Implementation of a specific protocol with a unique, centrally-regulated calibration procedure and clearly defined and pre-specified MRI positivity thresholds, minimizes interscanner variability45 and ensures precise and accurate results for use in clinical practice. In 2013 the U.S. Food and Drug Administration approved use of a T2 SE protocol (FerriScan®, Resonance Health Ltd) for estimating liver iron concentration as an imaging companion diagnostic for deferasirox therapy in patients with non-transfusion-dependent thalassemia.46 Nevertheless, there are only two studies assessing the diagnostic accuracy of this method in patients with iron overload,11, 35 hence it is crucial to accumulate more evidence to verify the diagnostic accuracy of this protocol. Moreover, future studies should ascertain the effect of MRI surveillance on patient survival and development of chronic liver disease. All in all, insufficient evidence depreciates the validity of clinical practice recommendations supporting routine use of MRI in patients with transfusional iron overload.7, 13, 14 It is thus imperative for guideline developers in the future to consider the totality of evidence, particularly because MRI is a costly intervention and its cost-effectiveness in this setting has not been established.
In conclusion, MRI may be accurate enough to rule out iron overload, and to rule in this condition particularly in individuals at high-risk. We draw this conclusion with low confidence in the estimates due to high risk of bias of the available studies. Our findings support the need for rigorous research that establishes the diagnostic accuracy of MRI, and evaluates the extent to which knowledge of the MRI test results leads to effective treatment that improves patients’ care and outcomes.

Acknowledgements

We are deeply grateful to authors of primary studies that responded to our request and provided additional data: Drs E. Alexopoulou, L. Anderson, J. Doornbos, Y. Gandon, M. Garbowski, J. Hankins, B. H. Kreeftenberg, S. Lawrence, T. St. Pierre, E. Rocchi, and J. Wood. We also thank Mr. Z. Wang for his valuable comments and feedback.
Figure 1. Flow diagram of study selection process.

Figure 2. Graphical display of Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) results. Methodological design and reporting quality of studies included in meta-analysis according to risk of bias and applicability concerns using the QUADAS-2 tool: review authors’ judgements about each methodological quality item presented as percentages across all included studies.

Figure 3. Forest plot of individual study estimates of sensitivity and specificity. Accuracy of Magnetic Resonance Imaging for estimating liver iron overload in patients with hereditary hemochromatosis or transfusional overload (thalassemia, sickle cell disease, myelodysplastic syndrome). Only studies reporting results for > 7 mg Fe/g dry liver weight biopsy threshold are included. FN, false negative; FP, false positive; TN, true negative; TP, true positive; CI, confidence interval.

Figure 4. Hierarchical summary receiver operating characteristic (HSROC) curve of sensitivity versus specificity. Performance of Magnetic Resonance Imaging for estimating liver iron concentration in patients with hereditary hemochromatosis or transfusional overload (thalassemia, sickle cell disease, myelodysplastic syndrome). Only studies reporting results for > 7 mg Fe/g dry liver weight biopsy threshold are included. Each circle represents a study, with the size being proportional to the study size. The solid curve represents the HSROC curve.

Figure 5. Conditional probability plots to explore the clinical utility of Magnetic Resonance Imaging. Clinical utility of (a) T2 spin echo and (b) T2* gradient recalled
echo sequences for detecting patients with liver iron overload (> 7 mg Fe/g dry liver weight) in all possible pre-test (prior) probabilities. LR, likelihood ratio; CI, confidence interval.

Supplementary Figure 1. Hierarchical summary receiver operating characteristic (HSROC) curve of sensitivity versus specificity for different Magnetic Resonance Imaging modalities. Performance of (a) T2 spin echo and (b) T2* gradient recalled echo sequences for estimating liver iron concentration in patients with hereditary hemochromatosis or transfusional overload (thalassemia, sickle cell disease, myelodysplastic syndrome). Each circle represents a study. The solid curve and respective solid points represent the HSROC curve and summary estimates of test performance. The dashed zone outlines surrounding them represent the 95% confidence region of these summary estimates.

Supplementary Figure 2. Hierarchical summary receiver operating characteristic (HSROC) curve of sensitivity versus specificity for different liver iron concentration thresholds. Performance of Magnetic Resonance Imaging for estimating liver iron concentration in patients with hereditary hemochromatosis or transfusional overload (thalassemia, sickle cell disease, myelodysplastic syndrome) in studies reporting results with a biopsy positivity threshold of 2 and 15 mg Fe/g dry liver weight (a and b respectively). Each circle represents a study, with the size being proportional to the study size. The solid curve represents the HSROC curve.
Supplementary Figure 3. Deeks’ funnel plot and asymmetry test for the detection of publication bias. Linear regression of the logistic diagnostic odds ratios against the inverse root of effective sample sizes (ESS).
References


2. van Bokhoven MA, van Deursen CT, Swinkels DW. Diagnosis and management of hereditary haemochromatosis. BMJ. 2011;342:c7251.


41. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994;271(9):703-707.


Table 1. Main characteristics of studies included in the systematic review and meta-analysis.

<table>
<thead>
<tr>
<th>Study, year</th>
<th>Population</th>
<th>Number of eligible patients</th>
<th>Mean age ± SD (range), years</th>
<th>Gender, % males</th>
<th>MRI sequence, measuring method</th>
<th>Time interval between MRI and biopsy ± SD (range), days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexopoulou et al, 2006</td>
<td>β-thalassemia</td>
<td>22</td>
<td>23 (12-34)</td>
<td>61</td>
<td>SE R2, relaxometry</td>
<td>45.4 (0-120)</td>
</tr>
<tr>
<td>Anderson et al, 2001</td>
<td>β-thalassemia</td>
<td>27</td>
<td>27.1 ± 6.7 (18-38)</td>
<td>60</td>
<td>GRE T2*, relaxometry</td>
<td>10 ± 7.0 (0-21)</td>
</tr>
<tr>
<td>Chan et al, 2001</td>
<td>Homozygous β-thalassemia</td>
<td>39</td>
<td>11.7 (3-22)</td>
<td>49</td>
<td>SE T1, SIR</td>
<td>&lt; 30</td>
</tr>
<tr>
<td>de Assis et al, 2012</td>
<td>β-thalassemia</td>
<td>11</td>
<td>(12-23)</td>
<td>43</td>
<td>GRE T2*, relaxometry</td>
<td>–</td>
</tr>
<tr>
<td>Engelhardt et al, 1994</td>
<td>Hemosiderosis, post-tranfusional hemosiderosis</td>
<td>13</td>
<td>39.9 (22-56)</td>
<td>77</td>
<td>SE T1 and SE T2, relaxometry</td>
<td>–</td>
</tr>
<tr>
<td>Gandon et al, 2004</td>
<td>Hemochromatosis</td>
<td>106</td>
<td>–</td>
<td>75</td>
<td>5 axial GRE sequences: T1, PD, T2, T2+, T2++, SIR</td>
<td>8 ± 23 (0-194)</td>
</tr>
<tr>
<td>Garbowska et al, 2009</td>
<td>Participants in iron chelation studies on deferasirox</td>
<td>18</td>
<td>–</td>
<td>–</td>
<td>GRE T2*,</td>
<td>&lt; 75</td>
</tr>
<tr>
<td>Hanksins et al, 2009</td>
<td>Sickle cell anemia, β-thalassemia, bone marrow failure syndromes</td>
<td>41</td>
<td>14 (7-35)</td>
<td>47</td>
<td>Multiecho GRE R2*, relaxometry</td>
<td>&lt; 30</td>
</tr>
<tr>
<td>Jensen et al, 2004</td>
<td>Hereditary hemochromatosis</td>
<td>8</td>
<td>(26-52)</td>
<td>45</td>
<td>SE T2, SIR</td>
<td>&lt; 14</td>
</tr>
<tr>
<td>Kreeftenberg et al, 2000</td>
<td>Suspected hereditary hemochromatosis</td>
<td>23</td>
<td>40</td>
<td>70</td>
<td>GRE, T1 SE and an intermediate pulse, SIR</td>
<td>&lt; 65</td>
</tr>
<tr>
<td>Lawrence et al, 1996</td>
<td>Hemochromatosis</td>
<td>43</td>
<td>55 (26-70)</td>
<td>72</td>
<td>SE T1 and PD SE, SIR</td>
<td>&lt; 30</td>
</tr>
<tr>
<td>Macfarlane et al, 1995</td>
<td>Idiopathic hemochromatosis</td>
<td>8</td>
<td>50 (31-63)</td>
<td>50</td>
<td>SE T2, relaxometry and SIR</td>
<td>15 (1-44)</td>
</tr>
<tr>
<td>Ooi et al, 2004</td>
<td>Homozygous β-thalassemia</td>
<td>32</td>
<td>18.5 ± 5.9</td>
<td>66</td>
<td>SE T1 and GRE T2, SIR</td>
<td>&lt; 180</td>
</tr>
<tr>
<td>Rocchi et al, 1993</td>
<td>Hemosiderosis, genetic hemochromatosis, myelodysplastic syndrome</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>SE T1 and SE T2, relaxometry</td>
<td>&lt; 365</td>
</tr>
<tr>
<td>Rose et al, 2006</td>
<td>Low-risk myelodysplastic syndrome, homozygous β-thalassemia</td>
<td>24</td>
<td>42 (23-79)</td>
<td>56</td>
<td>5 axial GRE sequences: T1, PD, T2, T2+, T2++, short echo time, SIR</td>
<td>–</td>
</tr>
<tr>
<td>St. Pierre et al, 2005</td>
<td>Hereditary hemochromatosis, thalassemia disorders, β-thalassemia/hemoglobin E</td>
<td>73</td>
<td>(8-74)</td>
<td>–</td>
<td>SE T2, relaxometry</td>
<td>&lt; 60</td>
</tr>
<tr>
<td>St. Pierre et al, 2013</td>
<td>β-thalassemia</td>
<td>233</td>
<td>14.3 ± 7.1 (3-43)</td>
<td>52</td>
<td>SE T2, relaxometry</td>
<td>15 ± 38</td>
</tr>
<tr>
<td>Tang et al, 2013</td>
<td>Transfusion-dependent thalassemia syndromes</td>
<td>8</td>
<td>(19-51)</td>
<td>50</td>
<td>Multiple SE R2 and R2*, relaxometry</td>
<td>180 (30-330)</td>
</tr>
<tr>
<td>Study</td>
<td>Disease(s)</td>
<td>N</td>
<td>Age (range)</td>
<td>Medication</td>
<td>Sequence/Relaxometry</td>
<td>T2* (SD)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------</td>
<td>-----</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Voskaridou et al, 2004</td>
<td>β-thalassemia, sickle cell disease</td>
<td>29</td>
<td>34.5 (19-65)</td>
<td>40</td>
<td>SE T2, relaxometry</td>
<td>–</td>
</tr>
<tr>
<td>Wood et al, 2005</td>
<td>Thalassemia major, sickle cell disease, thalassemia intermedia, aplastic anemia, hemochromatosis, heme-metabolism defect</td>
<td>21</td>
<td>–</td>
<td>–</td>
<td>GRE R2* and GRE R2, relaxometry</td>
<td>4.3 ± 9.5 (0-32)</td>
</tr>
</tbody>
</table>

SD, standard deviation; MRI, magnetic resonance imaging; SE, spin echo; R2, 1/T2; GRE, gradient recalled echo; T2*, transverse relaxation time which is affected by magnetic field inhomogeneity; T1, longitudinal relaxation time; SIR, signal intensity ratio; T2, transverse relaxation time; PD, proton density; R2*, 1/T2*.

* Data for the subset of participants analyzed (when available). Otherwise data for the total number of patients included in the respective study.
7,305 records identified through search of Medline (2,106), Embase (3,430), Cochrane Library (39) and Web of Science (1,730)

14 additional records identified from conference abstracts (13) and ClinicalTrials.gov (1)

2,482 duplicate records removed

4,837 records screened (title, abstract)

4,744 records excluded by screening title and abstract

93 full text records assessed for eligibility

73 full text records excluded:
15 beyond the scope of the review
12 not eligible population
19 not eligible index test
6 not eligible reference standard
21 a 2 x 2 table could not be reconstructed

20 studies included in systematic review and meta-analysis
<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Prevalence</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexopoulos et al., 2008</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>0.64</td>
<td>0.93 [0.88, 1.00]</td>
<td>1.00 [0.95, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anderson et al., 2001</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0.37</td>
<td>1.00 [0.63, 1.00]</td>
<td>0.53 [0.26, 0.77]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinn et al., 2001</td>
<td>12</td>
<td>4</td>
<td>6</td>
<td>17</td>
<td>0.46</td>
<td>0.67 [0.41, 0.87]</td>
<td>0.81 [0.56, 0.95]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de Assis et al., 2012</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0.64</td>
<td>0.86 [0.42, 1.00]</td>
<td>0.50 [0.07, 0.95]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gartovskii et al., 2009</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0.44</td>
<td>0.69 [0.47, 1.00]</td>
<td>0.90 [0.55, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawkins et al., 2006</td>
<td>29</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td>0.75</td>
<td>0.97 [0.83, 1.00]</td>
<td>0.91 [0.56, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jensen et al., 1994</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.75</td>
<td>1.00 [0.54, 1.00]</td>
<td>1.00 [0.16, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krettenburg et al., 2000</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>18</td>
<td>0.22</td>
<td>0.00 [0.00, 0.52]</td>
<td>1.00 [0.81, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lawrence et al., 1990</td>
<td>19</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0.73</td>
<td>1.00 [0.60, 1.00]</td>
<td>0.75 [0.54, 0.87]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macfarlane et al., 1995</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.75</td>
<td>0.67 [0.22, 0.99]</td>
<td>0.56 [0.01, 0.99]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olon et al., 2004</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td>14</td>
<td>0.48</td>
<td>0.86 [0.41, 0.89]</td>
<td>0.98 [0.62, 0.98]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosa et al., 2008</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>0.68</td>
<td>0.94 [0.70, 1.00]</td>
<td>1.00 [0.93, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. Pierre et al., 2003</td>
<td>59</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>0.77</td>
<td>0.69 [0.73, 0.86]</td>
<td>0.92 [0.57, 0.99]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. Pierre et al., 2013</td>
<td>167</td>
<td>4</td>
<td>15</td>
<td>47</td>
<td>0.79</td>
<td>0.92 [0.87, 0.99]</td>
<td>0.92 [0.81, 0.98]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tong et al., 2015</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0.03</td>
<td>Not estimable</td>
<td>Not estimable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voskardou et al., 2004</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>18</td>
<td>0.59</td>
<td>0.94 [0.71, 1.00]</td>
<td>0.95 [0.52, 0.98]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wood et al., 2005</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0.75</td>
<td>0.93 [0.68, 1.00]</td>
<td>1.00 [0.48, 1.00]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Positive test result
LR+ = 8.85 (95% CI 4.91, 15.93)

Negative test result
LR− = 0.10 (95% CI 0.07, 0.14)

Negative predictive value = 0.83 (95% CI 0.78, 0.88)
Positive predictive value = 0.81 (95% CI 0.77, 0.86)
Positive test result
LR+ = 4.86 (95% CI 1.74, 13.58)
Negative test result
LR− = 0.05 (95% CI 0.02, 0.17)

Negative predictive value = 0.88 (95% CI 0.79, 0.98)
Positive predictive value = 0.74 (95% CI 0.66, 0.83)
Search strategy for electronic databases

MEDLINE via PubMed

#1 "Iron overload" [mh]
#2 "iron overload" [tw]
#3 "iron/analysis" [mh]
#4 hemoglobinopathies [mh]
#5 "Hemoglobin, Sickle" [mh]
#6 hemoglobinopath*[tw]
#7 haemoglobinopath* [tw]
#8 thalassemia [tw]
#9 thalassaemia [tw]
#10 sickle cell [tw]
#11 drepanocytosis [tw]
#12 hemoglobin s [tw]
#13 haemoglobin s [tw]
#14 "hemoglobin sc disease" [tw]
#15 "haemoglobin sc disease" [tw]
#16 (sickling AND (blood OR plasma)) [tw]
#17 hemochromatosis [mh]
#18 hemochromatos* [tw]
#19 haemochromatos* [tw]
#20 myelodysplastic syndromes [mh]
#21 myeloplasti* [tw]
#22 myelodysplasti* [tw]
#23 dysmyelopoietic* [tw]
#24 myelodysplasi* [tw]
#25 mds [tw]
#26 OR/1-25
#27 "magnetic resonance imaging" [mh]
#28 "magnetic resonance imaging"[tw]
#29 mri [tw]
#30 MR imag* [tw]
#31 NMR imag* [tw]
#32 MR tomograph* [tw]
#33 "NMR tomograph*" [tw]
#34 OR/27-33
#35 34 AND 26

EMBASE via Ovid

#1 Iron overload.sh
#2 iron overload.mp
#3 hemoglobinopathy.sh.
#4 hemoglobinopath*.mp
#5 haemoglobinopath*.mp
#6 thalassemia.mp
#7 thalassaemia.mp
#8 sickle cell anemia.sh
#9 sickle cell.mp
#10 drepanocytosis.mp
#11 hemoglobin s.mp
#12 haemoglobin s.mp
#13 hemoglobin sc disease.mp
#14 haemoglobin sc disease.mp
#15 (sickling AND (blood OR plasma)).mp
#16 hemochromatosis.sh.
#17 hemochromatos*.mp
#18 haemochromatos*.mp
#19 myelodysplastic syndrome.sh.
#20 myeloplasti*.mp
#21 myelodysplasti*.mp
#22 dysmyelopoietic*.mp
#23 myelodysplasi*.mp
#24 mds.mp
#25 OR/1-24
#26 nuclear magnetic resonance imaging.sh.
#27 magnetic resonance imaging.mp.
#28 mri.mp
#29 magnetic resonance imag$.mp
#30 (MR adj2 imag$).mp.
#31 (NMR adj2 imag$).mp.
#32 (MR adj2 tomography).mp.
#33 (NMR adj2 tomography).mp.
#34 OR/26-33
#35 34 AND 25

Cochrane Library
#1 MeSH descriptor Iron overload explode all trees

#2 iron overload:ti,ab,kw

#3 MeSH descriptor hemoglobinopathies explode all trees

#4 MeSH descriptor Hemoglobin, Sickle explode all trees

#5 MeSH descriptor thalassemia explode all trees

#6 MeSH descriptor anemia sickle cell explode all trees

#7 hemoglobinopath*:ti,ab,kw

#8 haemoglobinopath*:ti,ab,kw

#9 thalassemi*:ti,ab,kw

#10 thalassaemi*:ti,ab,kw

#11 sickle next cell:ti,ab,kw

#12 sickled:ti,ab,kw

#13 sickling:ti,ab,kw

#14 drepanocytosis:ti,ab,kw

#15 (hemoglobin next s):ti,ab,kw

#16 (haemoglobin next s):ti,ab,kw

#17 (hemoglobin next sc next disease):ti,ab,kw

#18 (haemoglobin next sc next disease):ti,ab,kw

#19 (sickling and (blood or plasma)):ti,ab,kw

#20 MeSH descriptor Hemochromatosis explode all trees

#21 h?emochromatos*:ti,ab,kw

#22 MeSH descriptor Myelodysplastic Syndromes explode all trees

#23 myeloplasti*:ti,ab,kw

#24 myelodysplasti*:ti,ab,kw

#25 dysmyelopoietic*:ti,ab,kw
#26 myelodysplasi*:ti,ab,kw
#27 mds:ti,ab,kw
#28 OR/1-27
#29 MeSH descriptor magnetic resonance imaging explode all trees
#30 magnetic resonance imaging:ti,ab,kw
#31 mri:ti,ab,kw
#32 magnetic resonance imag*:ti,ab,kw
#33 MR adj2 imag*:ti,ab,kw
#34 NMR adj2 imag*:ti,ab,kw
#35 MR adj2 tomography:ti,ab,kw
#36 NMR adj2 tomography:ti,ab,kw
#37 OR/29-36
#38 37 AND 28

**Web of Knowledge**

#1 TS=(Iron overload)
#2 TS=(hemoglobinopathies)
#3 TS=(Hemoglobin, Sickle)
#4 TS=(hemoglobinopath*)
#5 TS=(haemoglobinopath*)
#6 TS=(thalassemia)
#7 TS=(thalassaemia)
#8 TS= (sickle cell)
#9 TS=(drepanocytosis)
#10 TS=(hemoglobin s)
#11 TS=(haemoglobin s)
#12 TS=(hemoglobin sc disease)
#13 TS=(haemoglobin sc disease)
#14 TS=(sickling AND (blood OR plasma))
#15 TS=(hemochromatosis)
#16 TS=(hemochromatos*)
#17 TS=(haemochromatos*)
#18 TS=(myelodysplastic syndrome)
#19 TS=(myeloplasti*)
#20 TS=(myelodysplasti*)
#21 TS=(dysmyeloipoietic*)
#22 TS=(myelodysplasi*)
#23 TS=(mds)
#24 OR/1-23
#25 TS=(magnetic resonance imaging)
#26 TS=(mri)
#27 TS=(MR imag*)
#28 TS=(NMR imag*)
#29 TS=(MR tomograph*)
#30 TS=(NMR tomograph*)
#31 OR/25-30
#32 31 AND 24
Hand-searched gray literature sources

- Annual scientific meetings of relevant associations:
  1. American Society of Hematology (from 2004 to 2012)
  2. European Hematology Association (from 2006 to 2013)
  3. Italian Society of Hematology (from 2007 to 2012)
  4. European Society of Radiology (from 2004 to 2013)
  5. European Society of Gastrointestinal and Abdomen Radiology (from 2004 to 2013)

- Web sites
  1. Thalassemia Reports – Open Access Journal
     (http://www.thalassemiareports.org/)
  2. FerriScan®, Resonance Health Ltd
     (http://www.resonancehealth.com/resonance/ferriscan)

- Trial registries (ClinicalTrials.gov)
Supplementary Figure 1
Supplementary Figure 2

a. Study estimate
   HSROC curve

b. Study estimate
   HSROC curve
Supplementary Figure 3
**Supplementary Table 1.** Diagnostic accuracy summary estimates for subgroups of studies using T2 spin echo or T2* gradient-recalled echo sequences.

<table>
<thead>
<tr>
<th></th>
<th>T2 spin echo</th>
<th>T2* gradient-recalled echo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Number of patients</td>
<td>393</td>
<td>118</td>
</tr>
<tr>
<td>Sensitivity (95% CI)</td>
<td>0.90 (0.85, 0.94)</td>
<td>0.96 (0.87, 0.99)</td>
</tr>
<tr>
<td>Specificity (95% CI)</td>
<td>0.87 (0.76, 0.93)</td>
<td>0.80 (0.53, 0.94)</td>
</tr>
<tr>
<td>Diagnostic odds ratio (95% CI)</td>
<td>59 (22, 158)</td>
<td>92 (15, 559)</td>
</tr>
<tr>
<td>Likelihood ratio of a positive test result (95% CI)</td>
<td>8.85 (4.91, 15.93)</td>
<td>4.86 (1.74, 13.56)</td>
</tr>
<tr>
<td>Likelihood ratio of a negative test result (95% CI)</td>
<td>0.10 (0.07, 0.14)</td>
<td>0.05 (0.02, 0.17)</td>
</tr>
</tbody>
</table>

T2, transverse relaxation time; T2*, transverse relaxation time which is affected by magnetic field inhomogeneity.