European Radiology (2020) 30:1959–1968
https://doi.org/10.1007/s00330-019-06450-y

MAGNETIC RESONANCE

# Impact of MRI technique on clinical decision-making in patients with liver iron overload: comparison of FerriScan- versus R2*-derived liver iron concentration

Marshall S. Sussman[1,2] · Richard Ward[3,4] · Kevin H. M. Kuo[3,4] · George Tomlinson[5,6] · Kartik S. Jhaveri[1,2]

Received: 23 May 2019 / Revised: 8 August 2019 / Accepted: 12 September 2019 / Published online: 17 January 2020


## Abstract

**Objectives** The purpose of this study was to compare clinical decision-making in iron overload patients using FerriScan and an R2*-based approach.

**Methods** One-hundred and six patients were imaged at two consecutive timepoints (454 ± 158 days) on a 1.5-T Siemens MAGNETOM Avanto Fit scanner. For both timepoints, patients underwent the standard FerriScan MRI protocol. During the second exam, each patient additionally underwent R2*-MRI mapping. For each patient, a retrospective (simulated) decision was made to increase, decrease, or maintain chelator levels. Two different decision models were considered: The fixed threshold model assumed that chelator adjustments are based strictly on fixed liver iron concentration (LIC) thresholds. Decisions made with this model depend only on the most recent LIC value and do not require any clinician input. The second model utilized decisions made by two hematologists retrospectively based on trends between two consecutive LIC values. Agreement ($\kappa_A$) between hematologists (i.e., interobserver variability) was compared with the agreement ($\kappa_B$) between a single hematologist using the two different LIC techniques.

**Results** Good agreement between R2*- and FerriScan-derived decisions was achieved for the fixed threshold model. True positive/negative rates were greater than 80%, and false positive/negative rates were less than 10%. ROC analysis yielded areas under the curve greater than 0.95. In the second model, the agreement in clinical decision-making for the two scenarios ($\kappa_A$ vs. $\kappa_B$) was equal at the 95% confidence level.

**Conclusions** Switching to R2*-based LIC estimation from FerriScan has the same level of agreement in patient management decisions as does switching from one hematologist to another.

**Key Points**

- *Good agreement between R2*- and FerriScan-derived decisions in liver iron overload patient management*
- *Switching to R2*-based LIC estimation from FerriScan has the same level of agreement in patient management decisions as does switching from one hematologist to another.*

**Keywords** Magnetic resonance imaging (MRI) · Iron · Liver · Decision-making · Uncertainty

✉ Kartik S. Jhaveri
kartik.jhaveri@uhn.ca

[1] Joint Department of Medical Imaging, University Health Network, Mount Sinai Hospital, and Women's College Hospital, University of Toronto, 610 University Ave, 3-957, Toronto, ON M5G 2M9, Canada

[2] Department of Medical Imaging, University of Toronto, Toronto, ON, Canada

[3] Division of Medical Oncology & Hematology, University Health Network, Toronto, ON, Canada

[4] Division of Hematology, University of Toronto, Toronto, ON, Canada

[5] Department of Medicine, University Health Network and Mount Sinai Hospital, Toronto, ON, Canada

[6] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

**Abbreviations**

| | |
|---|---|
| $\kappa$ | Kappa value |
| AUC | Area under the curve |
| LIC | Liver iron concentration |
| $LIC(R2*)$ | R2*-derived estimate of FerriScan LIC |
| $LIC_{\text{FerriScan}}$ | FerriScan-derived estimate of LIC |
| MRI | Magnetic resonance imaging |
| ROC | Receiver operating characteristic |

## Introduction

The assessment of liver iron concentration (LIC) is essential in the optimal management of patients with iron overload disorders [1]. Many patients, particularly those with anemia, are placed on iron chelation therapy. The historically and widely accepted target for LIC is between 3 and 7 mg iron/g liver (dry weight ≡ DW) [1, 2]; although with effective chelator choices now available, many are aiming to normalize LIC. A prolonged period of high LIC leads to iron overload complications (e.g., fibrosis, cirrhosis, and hepatocellular carcinoma), while a lower concentration could lead to chelator toxicity. A major factor in the clinical decision whether to adjust the chelator dose is the LIC level. An accurate assessment of LIC is therefore critical for patient management.

The current non-invasive method of choice for estimating LIC is MRI [3]. One of the most widely used MRI techniques is the FerriScan method, by Resonance Health, likely due to the fact that it has regulatory approval for providing MRI-derived LIC values. FerriScan generates an R2 map of the liver. It then applies a previously validated calibration curve to relate liver R2 values to iron concentration. However, FerriScan has a number of drawbacks: First, data must be transferred offsite for post-processing. This adds a delay to the receipt of information. Second, offsite processing incurs additional costs for each exam. Moreover, some researchers have argued for more frequent LIC testing in cases such as change in chelator regimen, rapid reduction in serum ferritin, or very high LIC [4]. The offsite processing costs present an impediment to this approach, especially from the public or private insurer's perspective. Finally, a complete FerriScan exam is relatively lengthy—typically at least 15 min.

As an alternative to FerriScan, a number of researchers have developed MRI-based R2* LIC quantification methods [5–10]. However, there has been some hesitation in adopting these methods due to the uncertainty in the relationship between LICs derived from R2* techniques and FerriScan's regulatory-approved values. There is concern that an alternative technique may lead to different clinical decisions relative to the FerriScan reference standard. In this paper, we explore this question in detail. Specifically, we analyze clinical decision-making using both FerriScan and an alternate R2*-based LIC quantification method that has been developed previously by our group [11]. The R2* scan itself can be completed in a breathhold.

## Methods

### MRI-based LIC quantification

All LIC data used in this paper were acquired as part of a previous study [11]. The purpose of that study was to investigate the sources of uncertainty between FerriScan- and R2*-derived LIC values. In that study, 106 patients with iron overload states due to hematological diseases were prospectively enrolled (mean age of 38 years, range of 19–72). The diseases included beta thalassemia (79), alpha thalassemia (4), sickle cell disease (7), hereditary spherocytosis (4), other chronic anemia (7), and other rare disorders (2). In three patients, there was no information about the disease type. These patients received FerriScan-MRI exams at two clinically indicated consecutive intervals (454 ± 158 days) on a 1.5-T Siemens MAGNETOM Avanto Fit scanner. The FerriScan data consisted of R2 maps of the liver. FerriScan converted these R2 maps into liver LIC values. During the second FerriScan exam visit, these patients additionally underwent R2* mapping MRI. R2* data were acquired using a prototype 3D bipolar 6-echo breathhold gradient echo acquisition [12] of the whole liver. Relevant pulse sequence parameters include the following: $TE_{min}$ = 1 ms, ΔTE = 1.4 ms, TR = 12 ms, flip angle 6°, $FOV_{read}$ = 400 mm, base matrix = 160, 48 slices of 4 mm thickness, parallel imaging acceleration × 3, and breathhold time 15 s.

The R2* data was converted in-house into LIC values ($\equiv \widehat{LIC(R2*)}$) using the equation derived in our previous study [11]:

$$\widehat{LIC(R2*)} = 0.0278 \cdot R_2^{*1.029} \quad (1)$$

In theory, LIC values derived from Eq. 1 could be used in lieu of FerriScan-derived LIC values to guide patient management. However, as demonstrated in our previous study, there is a range of uncertainty of up to 30% between FerriScan- and R2*-derived LIC values. This uncertainty could lead to discrepancies in clinical decision-making between the two techniques.

### Clinical decision-making

We analyzed the impact of using R2*-derived LIC values on the clinical decision of chelator dose adjustment using two separate decision models: The first model assumes that chelator decisions are strictly guided by fixed patient LIC thresholds. The decisions made with this model depend only on the most recent LIC value and do not require any clinician input.

Such a scenario may arise in the case of a patient's first or baseline MRI. The second model utilizes decisions made by expert hematologists in patient management based on trends between two consecutive LIC values [4]. Note that in all cases, patient management was simulated and retrospective. No actual clinical decisions were taken based on either model.

## Clinical decisions based on fixed thresholds

The objective of treatment in iron overload disorders is to maintain total body iron accumulation in a range that is as close to a normal as possible, while avoiding chelator toxicity. The historically and widely accepted target for LIC is between 3 and 7 mg iron/g liver (dry weight ≡ DW) [1, 2]. In a fixed threshold decision model, chelation treatment is altered if the LIC falls outside this range: chelation is increased if > 7, decreased if < 3 mg iron/g liver DW, and maintained at current levels otherwise. To simplify the analysis, decision-making for upper (7 mg iron/g liver DW) and lower (3 mg iron/g liver DW) thresholds will be considered separately.

To quantify the impact of using R2*-derived LIC for clinical decision-making, the true/false positive and true/false negative rates of the clinical decisions may be calculated:

$$\text{True Positive Rate} = \frac{\sum \text{True Positives}}{\sum \text{Condition Positives}} \equiv \text{Sensitivity} \quad (2)$$

$$\text{True Negative Rate} = \frac{\sum \text{True Negatives}}{\sum \text{Condition Negatives}} \equiv \text{Specificity} \quad (3)$$

$$\text{False Positive Rate} = \frac{\sum \text{False Positives}}{\sum \text{Condition Negatives}} \quad (4)$$

$$\text{False Negative Rate} = \frac{\sum \text{False Negatives}}{\sum \text{Condition Positives}} \quad (5)$$

A "condition positive" is defined as an LIC value which triggers that an action must be taken. For the purposes of this section, it will be assumed that clinical decisions based on FerriScan-derived LIC (i.e., $LIC_{\text{FerriScan}}$) are the reference standard. Therefore, for the upper threshold, a condition positive occurs if $LIC_{\text{FerriScan}}$ is above the threshold value (≥ 7 mg iron/g liver DW). In this case, an action (i.e., increased chelation) is required. A "condition negative" is defined as an LIC value that indicates no action is required. For the upper threshold, a condition negative occurs if the reference $LIC_{\text{FerriScan}}$ is below the threshold, i.e., < 7 mg/g DW. In this case, no action (i.e., no change in chelation) is required. For the case of the lower threshold, a condition positive occurs when $LIC_{\text{FerriScan}}$ is below the threshold (≤ 3 mg iron/g liver DW), since action is required (i.e., decrease chelation). A condition negative occurs when the $LIC_{\text{FerriScan}}$ is above the threshold. Figures 1 and 2 illustrate this concept.

We next assess the total number of true/false positives and true/false negatives (i.e., the numerators in Eqs. 2–5) associated with the use of $\widehat{LIC(R2^*)}$ for decision-making. We start by considering an upper $\widehat{LIC(R2^*)}$ threshold of 7 mg iron/g liver DW. The true/false positive and true/false negative regions may be identified in the manner shown in Fig. 1. As an example, the "false negative" region is where $LIC_{\text{FerriScan}} > 7$ mg iron/g liver DW, but $\widehat{LIC(R2^*)} \leq 7$ mg iron/g liver DW. For data in this region, the (incorrect) clinical decision based on $\widehat{LIC(R2^*)}$ would be for no change in chelation. The other regions may be analyzed similarly. The numerators in Eqs. 2–5 are calculated by summing over all of the data in each of the corresponding regions.

Figure 2 illustrates the true/false positive and true/false negative rates associated with a lower $\widehat{LIC(R2^*)}$ threshold of 3 mg iron/g liver DW.

In some clinical scenarios, it may be advantageous to alter the relative size of the true/false positive and true/false negative regions. This may be accomplished by altering the $\widehat{LIC(R2^*)}$ thresholds. Figure 3 illustrates an example where an upper $\widehat{LIC(R2^*)}$ threshold of 15 mg iron/g liver DW is used. With this threshold, the false positives have been completely eliminated, although this has come at the expense of decreased true positives and increased false negatives.

True/false positive and true/false negative rates were calculated using $\widehat{LIC(R2^*)}$ thresholds that covered the full range of values encountered in this study.

## Clinical decisions based on (simulated) patient management

In the fixed threshold model, the objective is to maintain the *current* LIC within a specified range (e.g., 3–7 mg iron/g liver DW). However, at many institutions, the decision to alter treatment is not based solely on the current LIC measurement [4]. Rather, clinical decisions are based in part on trends in LIC over time. For example, If LIC levels are > 7 mg iron/g liver DW, but decreasing on consecutive LIC assessments, the decision may be to maintain current chelation levels. To simulate patient management, anonymized LIC values from two consecutive FerriScan exams were presented to two hematologists for independent and blinded review. Each hematologist was required to make a recommendation as to increase, decrease, or maintain the current chelation. Subsequently, all FerriScan-derived LIC values at the second timepoint were replaced with R2*-derived LIC values. Clinical decision-making was then repeated in a similar manner.

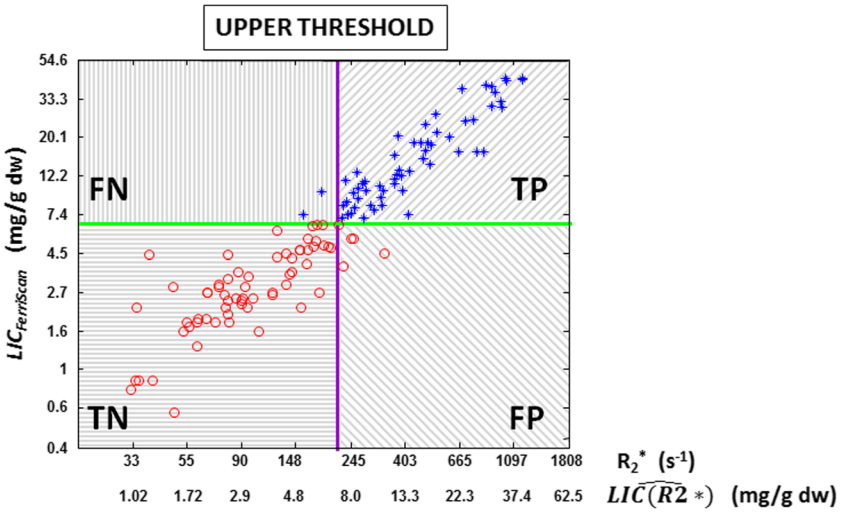


**Fig. 1** Decision-making for the upper LIC threshold (= 7 mg iron/g liver DW). The reference $LIC_{FerriScan}$ data are plotted against the corresponding R2* values, as well as the resulting maximum likelihood LIC estimates ($\widehat{LIC(R2*)}$) derived from Eq. 1. The green line indicates the 7-mg iron/g liver DW upper threshold value based on $LIC_{FerriScan}$. The "condition positive" data (blue asterisks) are those with $LIC_{FerriScan} > 7$ mg iron/g liver DW. If $LIC_{FerriScan}$ were used to guide treatment, these patients would receive an increase in chelation. The "condition negative" data (red circles) are those with $LIC_{FerriScan} \leq 7$ mg iron/g liver DW. If $LIC_{FerriScan}$ were used to guide treatment, these patients would receive no change in chelation. The purple line indicates an $\widehat{LIC(R2*)}$ value of 7 mg iron/g liver DW. If this value is used as a threshold for treatment (rather than $LIC_{FerriScan}$), some patients will be treated incorrectly. The four differently shaded rectangles indicate the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) regions

The patient order was randomized between the FerriScan and R2* analyses. Each reviewer was also blinded to the results of the other.

Clinical decisions were made under the following guidelines [4, 13]:

1) An LIC of < 5 mg/g DW was desired, and as close to normal range (1.18 mg/g DW) as possible.
2) For patients with very high LIC at baseline, a large reduction between the two scans would not trigger a dose change as chelator toxicity was not expected.

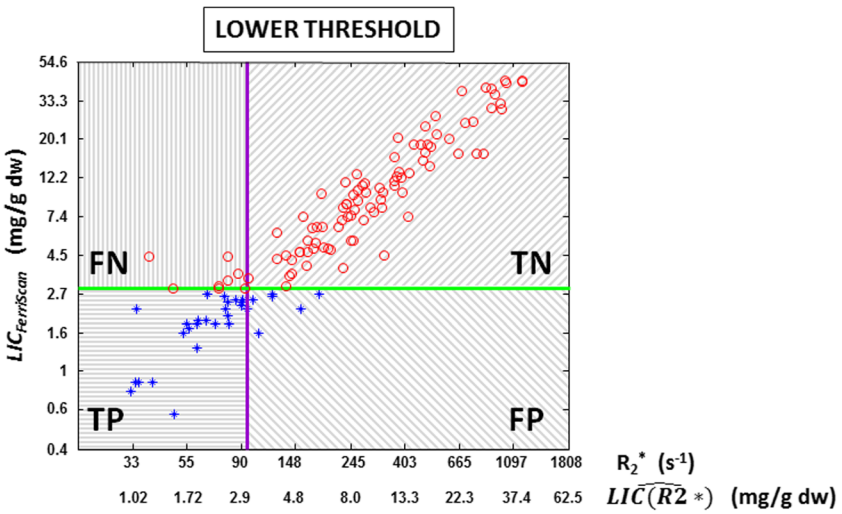


**Fig. 2** Decision-making for the lower LIC threshold (= 3 mg iron/g liver DW). The reference $LIC_{FerriScan}$ data are plotted against the corresponding R2* values, as well as the resulting maximum likelihood LIC estimates ($\widehat{LIC(R2*)}$) derived from Eq. 1. The green line indicates the 3-mg iron/g liver DW lower threshold value based on $LIC_{FerriScan}$. The "condition positive" data (blue asterisks) are those with $LIC_{FerriScan} \leq 3$ mg iron/g liver DW. If $LIC_{FerriScan}$ were used to guide treatment, these patients would receive a decrease in chelation. The "condition negative" data (red circles) are those with $LIC_{FerriScan} > 3$ mg iron/g liver DW. If $LIC_{FerriScan}$ were used to guide treatment, these patients would receive no change in chelation. The purple line indicates an $\widehat{LIC(R2*)}$ value of 3 mg iron/g liver DW. This R2* value corresponds to an LIC estimate of 3 mg iron/g liver DW. If this value is used as a threshold for treatment (rather than $LIC_{FerriScan}$), some patients will be treated incorrectly. The four differently shaded rectangles indicate the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) regions

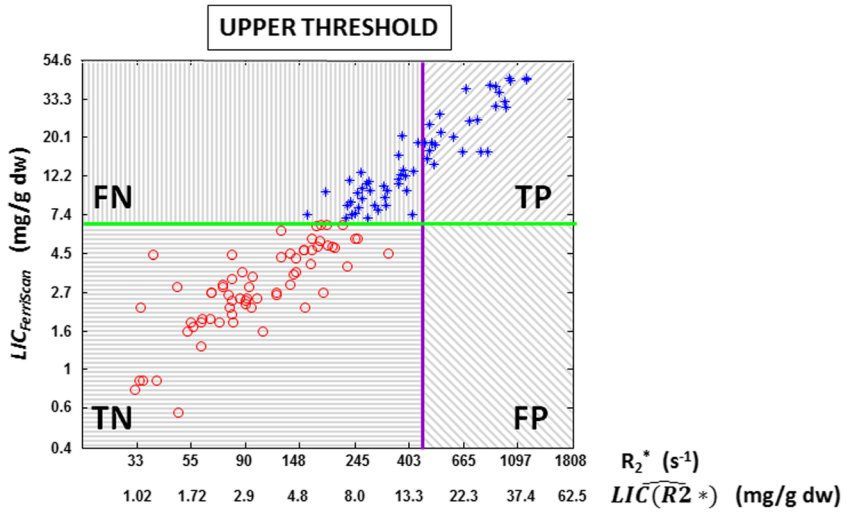


**Fig. 3** Decision-making for the upper LIC threshold (= 7 mg iron/g liver DW). The reference $LIC_{\text{FerriScan}}$ data are plotted against the corresponding R2* values, well as the resulting maximum likelihood LIC estimates ($\widehat{LIC(R2^*)}$ ) derived from Eq. 1. The green line indicates the 7-mg iron/g liver DW upper threshold value based on $LIC_{\text{FerriScan}}$. The "condition positive" data (blue asterisks) are those with true LIC > 7 mg iron/g liver DW. If $LIC_{\text{FerriScan}}$ were used to guide treatment, these patients would receive an increase in chelation. The "condition negative" data (red circles) are those with $LIC_{\text{FerriScan}} \leq 7$ mg iron/g liver DW. If $LIC_{\text{FerriScan}}$ were used to guide treatment, these patients would receive no change in chelation. The purple line indicates an $\widehat{LIC(R2^*)}$ value of 15 mg iron/g liver DW. If this value is used as a threshold for treatment (rather than $LIC_{\text{FerriScan}}$), some patients will be treated incorrectly. The four differently shaded rectangles indicate the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) regions

3) LICs in the low/low-medium range with a large reduction on follow-up would trigger a dose reduction ("soft landing"). This is due to concerns for impending chelator toxicity if the trend continued at the same trajectory.
4) Where the LIC was well within the normal range, and there was a further reduction on the second scan, a dose reduction was recommended for the same reasons as #3.
5) For very high LIC and with only a small reduction on the follow-up scan, a dose change was recommended due to suboptimal response.
6) Since chronically transfused patients continually load iron, stopping chelation was not an option.

The above protocol represents the (simplified) patient management workflow that is employed at our institution. It provides guidelines, rather than a rigid algorithm for patient management. As a consequence, patient management depends in part on the judgment of the individual hematologist. Therefore, there may be differences in patient management between hematologists (i.e., interobserver variability). The objective in this section is to determine whether differences in clinical decision-making caused by replacing FerriScan with R2* are significant in comparison with interobserver variability.

## Statistical methods used for clinical decision-making

The objective of the statistical analysis is to determine whether R2*-derived LIC values may be used as an acceptable alternative to FerriScan for clinical decision-making in iron overload disease. To make this determination, we consider two different scenarios (Table 1). In scenario A, we consider the level of agreement between two hematologists assessing the same FerriScan data sets (i.e., interobserver variability). Since this represents the situation occurring in actual clinical practice, it will be assumed that the level of agreement achieved in this scenario is clinically acceptable. In scenario B, we consider the level of agreement between the *same* hematologist making clinical decisions using R2*- versus FerriScan-derived LIC values. If this agreement is similar to the (clinically acceptable) level of agreement found in scenario A, then, R2* may be considered an option for clinical decision-making in iron overload disease.

To assess the levels of agreement discussed above, kappa coefficients are calculated. First, we calculate the kappa coefficient between management decisions under scenario A ($\equiv \kappa_A$) and for each reader under scenario B ($\equiv \kappa_B$). We then estimate the difference between $\kappa_A$ and $\kappa_B$ and its 95% confidence interval; a narrow interval is evidence that the level of agreement between scenarios A and B is similar. We also compute a $p$ value to test the hypothesis that the difference between $\kappa_A$ and $\kappa_B$ is 0. Since both kappa estimates use some of the same data, they are not independent. As a result, no closed form solution exists for inferences on this difference of kappas. Instead, we use the blocked bootstrap method [14]. In this approach, the entire set of readings for each patient subject is a block that is resampled using the non-parametric bootstrap. Kappa values and their differences are calculated on each bootstrap sample. The distribution of these values across 2000 bootstrap samples is used to (a) calculate a percentile-based 95%

**Table 1** Two different scenarios compared for clinical decision-making. In scenario A, decisions are made based on a comparison of FerriScan-derived LIC values at both timepoints. Here, the agreement ($\kappa_A$) is calculated between the decisions made by two hematologists. This scenario is representative of current clinical practice. In scenario B, we determine the agreement ($\kappa_B$) between the clinical decisions made by a single hematologist when the first decision is based on two FerriScan-derived LIC timepoints, and the second decision is based on a comparison of an initial FerriScan- and a second R2*-derived LIC timepoint. Although not indicated in the table, the scenario B analysis is also performed by hematologist 2

| | Timepoint #1 | Timepoint #2 | Decision-makers | Agreement between decisions |
|---|---|---|---|---|
| **Scenario A (current clinical practice)** | FerriScan | FerriScan | Hematologist 1 | $\kappa_A$ |
| | FerriScan | FerriScan | Hematologist 2 | |
| **Scenario B** | FerriScan | FerriScan | Hematologist 1 | $\kappa_B$ |
| | FerriScan | R2* | Hematologist 1 | |

confidence interval for the difference and (b) calculate the standard error of the difference, which can be used in a $z$-test with the observed difference to calculate the approximate significance level from which the hypothesis may be tested.

## Results

For the fixed threshold model, Table 2 lists the true/false positive and true/false negative rates for upper and lower $\widehat{LIC(R2^*)}$ thresholds of 7 and 3 mg iron/g liver DW respectively. Figure 4a plots the true/false positive and true/false negative rates over the full range of $\widehat{LIC(R2^*)}$ thresholds. For reference, the solid orange line indicates the rates achieved when an $\widehat{LIC(R2^*)}$ threshold of 7 iron/g liver DW is employed (note that the intersection of the orange line and the various curves corresponds to the data listed in Table 2). If a different $\widehat{LIC(R2^*)}$ threshold is used, then, the true/false positive and true/false negative rates (relative to the FerriScan-derived LIC reference standard) will change. Figure 4b illustrates this concept as an ROC curve. The area under the curve (AUC) is 0.97. Figure 5 is the same plot for the lower threshold data. The AUC is 0.95.

For the simulated patient management model, Table 3 lists the data for scenario A. The percent agreement is 82%, and $\kappa_A$ is 0.79. As discussed previously, since this represents current practice, we take this level of agreement as clinically acceptable. Table 4 lists the data for scenario B. The percent agreement and $\kappa_B$ for reader 1 are 82.4% and 0.76 respectively. The percent agreement and $\kappa_B$ for reader 2 are 85% and 0.81 respectively. Qualitatively, these results indicate that the level of agreement between scenarios A and B is very similar. Quantitatively, Table 5 shows the results from the bootstrap analysis assessing $\kappa_A - \kappa_B$. The $p$ values are large, providing no evidence of differences between the two kappa values. This implies that the level of agreement in switching to R2*- versus FerriScan-derived LIC measures (scenario B) is about the same as the level of agreement associated with different clinicians assessing the same data (scenario A).

## Discussion

The management of patients with iron overload states is multifactorial and complex. Clinical decision-making regarding adjustment of chelation therapy includes many considerations beyond iron overload status [15–18]. These include the following: chelation drug and dosage, transfusion and chelation therapy side effects (e.g., arrhythmias, cardiomyopathies, diabetes, renal dysfunction, pubertal hypogonadism, and growth retardation), and laboratory data (e.g., serum ferritin level and ferritin trend, proteinuria, serum creatinine, ALT, absolute neutrophil count). Despite these additional considerations, quantification and serial monitoring of the degree of iron burden (represented by liver iron concentration) is fundamental to the management of patients with iron overload. This study therefore attempted to address the question as to whether the use of different MRI-based LIC measurement strategies has a significant impact on iron chelation management decisions.

In the fixed threshold model, clinical decisions are based on whether the LIC is outside of the desired range (in this study, 3–7 mg iron/g liver DW). Decision-making was analyzed in terms of the true/false positive and true/false negative rates. Overall, there was very good agreement between

**Table 2** True/false positive and negative rates for upper and lower R2*-derived LIC ( $= \widehat{LIC(R2^*)}$ ) thresholds in the fixed decision model

| $\widehat{LIC(R2^*)}$ threshold | True positive rate | True negative rate | False positive rate | False negative rate |
|---|---|---|---|---|
| Upper (7 mg iron/g liver DW) | 0.95 | 0.94 | 0.06 | 0.05 |
| Lower (3 mg iron/g liver DW) | 0.81 | 0.91 | 0.09 | 0.19 |

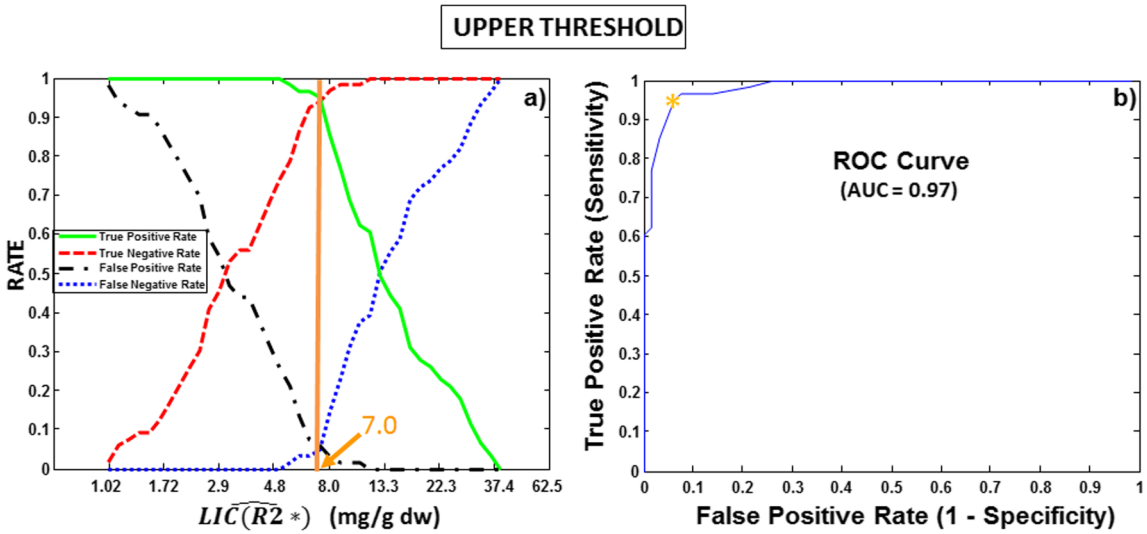


**Fig. 4** Upper threshold data for (**a**) true/false positive and true/false negative rates. The orange line indicates an $\widehat{LIC(R2*)}$ threshold of 7 mg iron/g. **b)** ROC curve, with area under the curve (AUC) indicated. The orange asterisk corresponds to the 7-mg iron/g threshold

FerriScan- and R2*-based decision-making (AUC > 95%). However, there were some cases where discrepancies occurred. In the context of iron overload treatment, the decisions that most adversely affect patient outcome are false negatives. If a false negative occurs, patients who have an LIC outside the desired range will (incorrectly) not receive a change in chelation. In these cases, iron levels would remain either too high (if the false negative is related to upper threshold) contributing to continued liver injury and complications of iron overload, or too low (if related to the lower threshold) and place the patient at risk of chelator toxicity. Thus, one strategy could be to set thresholds that minimize false negatives. On the other hand, the false positive rate must also be considered. These patients have FerriScan-derived LICs in the appropriate range, but R2*-derived LIC values would indicate that a change in chelation is required. The results of this study may be used to establish R2*-derived LIC thresholds that strike an appropriate balance over the patient population. For example, Figure 5b indicates that a substantial increase in the true positive rate could be achieved if the 3-mg iron/g liver DW lower threshold is adjusted to permit a slight increase in the false positive rate.

The second type of clinical decision-making analyzed in this study was based on the simulated management of patients in our study. The results indicated that agreement between different reviewers assessing the same FerriScan data (i.e., *interobserver* variability) was the same magnitude as the agreement when same reviewer used R2*- compared with

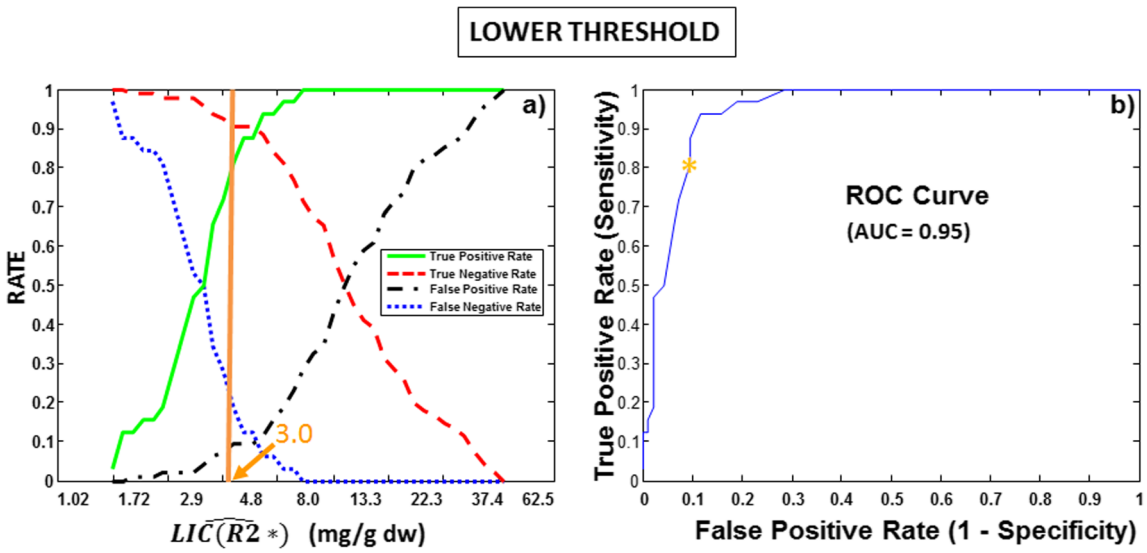


**Fig. 5** Lower threshold data for (**a**) true/false positive and true/false negative rates. The orange line indicates an $\widehat{LIC(R2*)}$ threshold of 3 mg iron/g. **b** ROC curve, with area under the curve (AUC) indicated. The orange asterisk corresponds to the 3-mg iron/g threshold

**Table 3** Agreement between readers (scenario A) in the simulated clinical management model. The decision to decrease (*d*), maintain (*m*), or increase (*i*) chelation for each read is listed. The percent agreement is 82.4%, and $\kappa_A$ is 0.79 (95% CI 0.70–0.89)

| | | Hematologist 1 | | |
|---|---|---|---|---|
| | | *d* | *m* | *i* |
| **Hematologist 2** | *d* | 9 | 2 | 0 |
| | *m* | 0 | 37 | 23 |
| | *i* | 0 | 1 | 40 |

FerriScan-derived LIC values. Thus, from a statistical point of view, switching to R2*-based decision-making from FerriScan is no more likely to cause a difference in patient management than switching decision-making from one hematologist to another. It is also worth examining the nature of the disagreement in the two scenarios. In the case of interobserver variability (scenario A), the two hematologists disagreed mostly on whether there should be no change versus an increase in chelation. In contrast, the disagreement in scenario B (FerriScan- vs. R2*-based decision-making) was more uniformly distributed among the different treatment possibilities. However, there were no cases in either scenario where increase in chelation by one method or hematologist corresponded to a decrease in chelation by the other method or hematologist (or vice versa). These latter situations represent the most severe (and problematic) form of disagreement.

Other works comparing clinical decision-making between R2* and FerriScan techniques include a study by Chan et al [5]. They found a sensitivity and specificity of 95% and 89% respectively for the detection of LIC > 7 mg iron/g liver DW by R2* (using FerriScan as the reference standard). In contrast, our fixed threshold model had a sensitivity and specificity of 95% and 94% respectively for R2* (see Fig. 4b). A study by Nichols-Venezula et al [19] found a false positive rate of about 14% and a false negative rate of 1% for the detection of LIC > 7 mg iron/g liver DW by R2* (again, with FerriScan as the reference standard). For comparison, our study had a false positive rate of 6% and a false negative rate of 5%.

**Table 4** Agreement between clinical decisions made with FerriScan- and R2*-derived LIC values (scenario B) for hematologists 1 and 2 in the simulated clinical management model. The number of patients with decisions to decrease (*d*), maintain (*m*), or increase (*i*) chelation is listed. For hematologist 1, the percent agreement is 82.4%, and $\kappa_B$ is 0.76 (95% CI 0.67–0.86). For hematologist 2, the percent agreement is 85.2%, and $\kappa_B$ is 0.81 (95% CI 0.72–0.90)

| Hematologist | | 1 | | | 2 | | |
|---|---|---|---|---|---|---|---|
| | | FerriScan | | | FerriScan | | |
| | | *d* | *m* | *i* | *d* | *m* | *i* |
| **R2*** | *d* | 5 | 2 | 0 | 6 | 1 | 0 |
| | *m* | 6 | 49 | 9 | 3 | 32 | 5 |
| | *i* | 0 | 2 | 35 | 0 | 7 | 54 |

The clinical decision models used in this study assumed that FerriScan provides true, gold standard measures of LIC. In turn, this implied that any discrepancy between R2*- and FerriScan-derived LIC values was due to deficiencies in the R2* technique. In reality, this assumption is likely overly pessimistic. The dominant source of discrepancy between FerriScan-derived LIC and the LIC values obtained using our R2* protocol was found to be the spatial heterogeneity of iron across the liver [11]. This was because the two techniques analyzed different regions of the liver. However, there is (to our knowledge) no study which demonstrates the "optimal" liver ROI for LIC assessment. It is therefore not clear that decisions made based on R2* would be necessarily inferior to those guided by FerriScan. Note that if the R2* liver regions were more closely matched to those of FerriScan, it is likely that the discrepancy in LIC values (and thus clinical decision-making) would be correspondingly reduced. However, the discrepancy would likely not be completely eliminated. Less prominent sources of variation (e.g., differences in MRI acquisition parameters, relaxometry methodology, or physical modes of iron relaxation) would remain [8].

There are a few limitations of this study that should be mentioned: Firstly, all statistics were calculated based on the patient population at our institution. The patient mix and, in particular, the distribution of LICs may differ at other sites. In turn, this may impact some of the results—especially the true/false positive and true/false negative calculations. On the other hand, our institution is the largest thalassemia center in North America. Therefore, most patient types are likely represented at our site. Secondly, the analysis of clinical decision-making was performed using the current Canadian consensus standard for patient management [4]. If the strategies used by other institutions deviate significantly from this, the results of this study may not be directly applicable. However, note that US guidelines are similar [13] . Thirdly, the results of this study cannot necessarily be extrapolated to the pediatric population. Though the principles of adjusting chelation are similar, the degree of tolerable iron overloading as well as the need to not overchelate may be different in children. Finally, as mentioned previously, the patient management protocol used in this study was somewhat simplified in comparison with actual patient management used in the clinic. Our intent was to isolate the impact of R2* and FerriScan on clinical decision-making from other confounding variables. In particular, in addition to LIC values, patient management is individualized and takes into account other markers of iron overload morbidity as well as patient-specific iron chelator medication adherence factors. It was also assumed that patients fully adhered to their chelation, the chelator dose was not already at maximum, there was no chelator toxicity, and there was no

**Table 5** Comparison of difference in kappa values in scenarios A and B in the simulated clinical management model. The null hypothesis being tested is that the two kappa values are equal (i.e., there is no difference in the level of agreement between scenarios A and B)

| Hematologist | $\kappa_A$ | $\kappa_B$ | $\kappa_A-\kappa_B$ | 95% confidence level for $\kappa_A-\kappa_B$ | | Approximate $p$ value |
|---|---|---|---|---|---|---|
| | | | | Lower | Upper | |
| 1 | 0.79 | 0.76 | 0.03 | − 0.10 | 0.15 | 0.65 |
| 2 | | 0.81 | − 0.02 | − 0.13 | 0.09 | 0.73 |

significant cardiac siderosis. The above assumptions are not fully consistent with real-life practice. However, they were employed to allow the clinicians to focus specifically on the role of LIC values in decision-making.

One issue that should be mentioned is the fact that R2* technique for LIC estimation was utilized along with FerriScan only at the second timepoint (FerriScan alone was used at the first timepoint). This was done in order to minimize the number of variables, and therefore minimize the random uncertainty (which, from our previous study [11], is known to be large). However, one problem with this approach is that if there is significant systematic error (e.g., a bias or offset) between FerriScan and R2*, this could in turn bias clinical decision-making when using different techniques at the two timepoints. Fortunately, our previous study demonstrated that the uncertainty between LICs derived from FerriScan and R2* is predominantly random [11]. Furthermore, the results of this study show that, even if such a bias in clinical decision-making exists, it is at a level that is not significant in comparison with interobserver variability. A more comprehensive study would compare clinical decision-making using FerriScan and R2* over an extended time period. However, note that our previous study [11] showed that much of the R2*-FerriScan discrepancy is likely caused by differences in liver ROIs analyzed between the two techniques. Therefore, to properly interpret the results, it would also be necessary to compare clinical decision-making using FerriScan alone with different ROIs.

The study demonstrated that clinical decision-making in iron overload disease may be affected by switching from FerriScan to R2*. It also showed that decision-making may be impacted by interobserver variability. In our study, we demonstrated that the magnitude of agreement is similar in both situations. However, the more general question as to the ultimate clinical impact of this variability on patient outcome is at present unknown. This question was not addressed in our study, nor (to the best of our knowledge) in any other study to date. To properly investigate the actual real-life impact, a long-term observational cohort study would be required assessing impact of each technique and other confounding factors on eventual patient outcome measures (e.g., cardiac complications, development of endocrine disorders, liver cirrhosis and failure, and death).

The objective of this study was to examine the impact on clinical decision-making in iron overload patients using FerriScan versus an R2*-derived LIC approach. Our study addressed this question by examining two different clinical decision-making models between FerriScan and an R2*-based technique: In the fixed threshold model, true/false positive/negative rates and ROC curves were calculated. Good agreement was achieved, with AUC > 0.95. In the patient management model, the agreement in clinical decision-making between R2*- and FerriScan-derived LICs was found to be similar to the agreement associated with interobserver variability using FerriScan alone. This implied that switching to R2*-based LIC estimation from FerriScan has the same level of agreement in patient management decisions as does switching from one hematologist to another.

**Funding information** The authors state that this work has not received any funding.

## Compliance with ethical standards

**Guarantor** The scientific guarantor of this publication is Kartik Jhaveri.

**Conflict of interest** The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

**Statistics and biometry** One of the authors has significant statistical expertise.

**Informed consent** Written informed consent was obtained from all subjects (patients) in this study.

**Ethical approval** Institutional Review Board approval was obtained.

**Study subjects or cohorts overlap** All MRI data used in this manuscript were acquired as part of a previous study (Jhaveri et al, JMRI 2018, pp. 1467–1474). The purpose of this previous study was twofold: first, to define a calibration curve that related R2*-derived liver iron concentration (LIC) values to FerriScan-derived LIC values; second, to characterize the nature of the uncertainty in the relationship between those two quantities. In the present manuscript, (simulated) clinical decisions were made using the R2*- and FerriScan-derived LIC values respectively. An analysis was performed to determine the agreement in clinical decision-making between these two approaches. Although the same data is used for both studies, there is no overlap in the nature or scope of the work.

**Methodology**
• diagnostic or prognostic study
• performed at one institution

## References


1. Olivieri NF, Brittenham GM (1997) Iron-chelating therapy and the treatment of thalassemia. Blood 89:739–761
2. Ferriscan (2015) MRI measurement of liver iron concentration. Available via http://www.resonancehealth.com/images/files/FerriScan/FerriScan%20Fact%20Sheet%20Mar%202015.pdf
3. St Pierre TG, Clark PR, Chua-Anusorn W et al (2005) Noninvasive measurement and imaging of liver iron concentrations using proton magnetic resonance. Blood 105:855–861
4. Mark Belletrutti LB, Corriveau-Bourque C, Ezzat H et al (2018) Consensus statement of clinical care of patients with thalassemia in Canada. Available via https://www.canhaem.org/wp-content/uploads/2018/10/consensus-statment-Thalassemia-Final.pdf
5. Chan WC, Tejani Z, Budhani F, Massey C, Haider MA (2014) R2* as a surrogate measure of FerriScan iron quantification in thalassemia. J Magn Reson Imaging 39:1007–1011
6. Garbowski MW, Carpenter JP, Smith G et al (2014) Biopsy-based calibration of T2*magnetic resonance for estimation of liver iron concentration and comparison with R2 FerriScan. J Cardiovasc Magn Reson 16:40
7. Runge JH, Akkerman EM, Troelstra MA, Nederveen AJ, Beuers U, Stoker J (2016) Comparison of clinical MRI liver iron content measurements using signal intensity ratios, R (2) and R (2)*. Abdom Radiol (NY) 41:2123–2131
8. Wood JC, Enriquez C, Ghugre N et al (2005) MRI R2 and R2* mapping accurately estimates hepatic iron concentration in transfusion-dependent thalassemia and sickle cell disease patients. Blood 106:1460–1465
9. Hankins JS, McCarville MB, Loeffler RB et al (2009) R2* magnetic resonance imaging of the liver in patients with iron overload. Blood 113:4853–4855
10. Henninger B, Zoller H, Rauch S et al (2015) R2* relaxometry for the quantification of hepatic iron overload: biopsy-based calibration and comparison with the literature. Rofo 187:472–479
11. Jhaveri KS, Kannengiesser SAR, Ward R, Kuo K, Sussman MS (2019) Prospective evaluation of an R2*method for assessing liver iron concentration (LIC) against FerriScan: derivation of the calibration curve and characterization of the nature and source of uncertainty in the relationship. J Magn Reson Imaging 49:1467–1474
12. Zhong XD, Nickel MD, Kannengiesser SAR, Dale BM, Kiefer B, Bashir MR (2014) Liver fat quantification using a multi-step adaptive fitting approach with multi-echo GRE imaging. Magn Reson Med 72:1353–1365
13. Sheth S (2018) Monitoring of iron overload in transfusion-dependent thalassemia (TDT). New York. Available via http://www.thalassemia.org/boduw/wp-content/uploads/2018/05/Monitoring-of-Iron-Overload-in-Transfusion-Dependent-Thalassemia-TDT-1.pdf
14. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
15. Tanhehco YC, Shi PA (2019) Chapter 52 - Transfusion management of patients with sickle cell disease and thalassemia. In: Shaz BH, Hillyer CD, Reyes Gil M (eds) Transfusion medicine and hemostasis (third edition). Elsevier, pp 325–335
16. Papakonstantinou O, Ladis V, Kostaridou S et al (2007) The pancreas in beta-thalassemia major: MR imaging features and correlation with iron stores and glucose disturbances. Eur Radiol 17:1535–1543
17. Lekawanvijit S, Chattipakorn N (2009) Iron overload thalassemic cardiomyopathy: Iron status assessment and mechanisms of mechanical and electrical disturbance due to iron toxicity. Can J Cardiol 25:213–218
18. Borgna-Pignatti C, Rugolotto S, De Stefano P et al (2004) Survival and complications in patients with thalassemia major treated with transfusion and deferoxamine. Haematologica 89:1187–1193
19. Nichols-Vinueza DX, White MT, Powell AJ, Banka P, Neufeld EJ (2014) MRI guided iron assessment and oral chelator use improve iron status in thalassemia major patients. Am J Hematol 89:684–688