Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.mrijournal.com

Liver MRI is more precise than liver biopsy for assessing total body iron balance: a comparison of MRI relaxometry with simulated liver biopsy results[☆],[☆☆]

John C. Wood [a,*], Pinggao Zhang [b], Hugh Rienhoff [d], Walid Abi-Saab [c], Ellis J. Neufeld [e]

[a] *Department of Pediatrics and Radiology, Children's Hospital Los Angeles*
[b] *Shire Pharmaceuticals, Chesterbrook, PA*
[c] *Shire, Global Development, Eysins, Switzerland*
[d] *Ferrokin Biosciences Inc., San Carlo, CA*
[e] *Boston Children's Hospital and Harvard Medical School, Boston, MA*

ARTICLE INFO

Article history:
Received 15 September 2014
Revised 14 January 2015
Accepted 16 February 2015
Available online xxxx

Keywords:
Relaxometry
Iron overload
Liver
Thalassemia
Sickle cell disease
Monte-Carlo Simulation

ABSTRACT

Purpose: Liver biopsy was long considered the reference standard for measuring liver iron concentration. However, its high sampling variability and invasive nature make it poorly suited for serial analyses. To demonstrate the fallibility of liver biopsy, we use serial estimates of iron chelation efficiency (ICE) calculated by R2 and R2* MRI liver iron concentration (LIC) estimates as well as by simulated liver biopsy (over all physically reasonable sampling variability) to compare the robustness of these three techniques.
Materials and Methods: R2, R2*, transfusional volume, and chelator compliance were obtained from 49 participants in a phase II clinical trial of deferitazole over two years. Liver biopsy LIC results were simulated using sampling errors of 0%, 10%, 20%, 30%, 40% and iron assay variability of 12%. LIC estimates by R2, R2*, and simulated biopsy were used to calculate ICE over time. Bland–Altman limits of agreement were compared across observation intervals of 12, 24, and 48 weeks.
Results: At 48 week intervals, LIC estimates by R2, R2* and "perfect" liver biopsy had comparable accuracy in predicting ICE; both MRI methods were superior to any physically realizable liver biopsy (sampling error 10% or higher). LIC by R2* demonstrated the most robust ICE estimates at monitoring intervals of 24 and 12 weeks, but this difference did not remain significant at 48 week intervals.
Conclusion: MRI relaxometry is superior to liver biopsy for serial LIC observations, such as used in the care of tranfusional siderosis patients, and should also be considered the new standard of LIC determination for regulatory purposes. Among relaxometry techniques, LIC estimates by R2* are more robust for tracking changes in iron balance over intermediate time scales (<=24 weeks).

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Patients on chronic blood transfusion require lifelong iron chelation therapy and monitoring of iron stores [1,2]. Ferritin levels are often used to guide iron chelation therapy, but there is large inter-subject variability between ferritin levels and iron stores because of confounding influences from systemic inflammation, ascorbate status, and other factors. In contrast, liver iron concentration LIC is a robust metric of iron balance [3] and total body iron stores [4] in transfused patients. Liver biopsy was once routinely performed for this purpose in thalassemia patients [5] but suffers from high sampling variability [6–8], processing biases [9], 0.5% major complication rate [10], and poor patient acceptance. As a result, it is gradually being replaced by MRI relaxometry methods in clinical care [11]. At least four R2* analysis software packages have received FDA approval, and one laboratory (Resonance Health) has received 510 K certification for its Ferriscan® R2 acquisition and analysis techniques. Nonetheless, the FDA and other regulatory authorities have been reluctant to accept MRI relaxometry as a surrogate for LIC in phase 3 clinical trials, despite multiple, independent cross-sectional studies validating both R2 [12,13] and R2* LIC estimation [13,14].

Since neither liver biopsy, nor MRI relaxometry can be considered a "gold standard", there is need for an independent mechanism to assess their accuracy and reproducibility. We propose that the

[☆] This work was supported by Shire Pharmaceuticals and the National Institute of Health (1R01DK097115-01A1).

[☆☆] Data from this study were presented as a poster at the American Society of Hematology Meetings, New Orleans, Dec 7–10, 2013.

[*] Corresponding author at: Children's Hospital of Los Angeles, Division of Cardiology, MS #34, 4650 Sunset Blvd, Los Angeles, CA, 90027.

http://dx.doi.org/10.1016/j.mri.2015.02.016
0730-725X/© 2015 Elsevier Inc. All rights reserved.

variance in longitudinal estimates of iron chelation efficiency (ICE) represents a physiologically relevant performance metric of any LIC estimator. ICE is defined as the molar ratio of the iron eliminated from the body to the maximum iron binding capacity of the consumed drug [15]. It can be estimated by knowledge of blood transfusion volumes and iron content, actual drug consumed and changes in LIC; no a priori knowledge drug bioavailability is needed. ICE varies widely across patients, but should remain relatively constant for any given patient as long as transfusion and chelation practices are stable, and there is no change in the hepatitis C status. In practice, ICE is generally less than 50% because of slow kinetics of the chelatable iron pool. If the LIC estimates were perfect, the resulting ICE estimates would have values between 0 and 1 and vary little over time. However, LIC measurement errors translate into inconsistent and potentially nonsensical estimates of ICE [i.e. values less than zero or greater than 1]. In this manuscript, we use this ICE concept to examine the performance of R2 and R2* LIC assessments and compare them to theoretical performance that would have been predicted by liver biopsy across a broad range of sampling variability estimates.

2. Materials and methods

All patient studies were performed as part of a phase II clinical trial of deferitazole (formerly FBS0701 or SPD602) that has been previously reported [16]. All participants provided informed consent, and study was performed according to Good Clinical Practice and HIPAA compliance under an IRB approved protocol. Patients were required to have baseline LIC by R2 less than 30 mg/g dry weight. The primary study and its 96 week extension phase are registered at ClinicalTrials.gov (NCT01186419, NCT0167111). 49 patients completed 24 weeks of therapy, 39 completed 48 weeks of therapy, and 26 completed the two-year extension study. Patient's ages ranged from 18 to 60 (28.5 ± 7.8 years). Most patients had a thalassemia syndrome (38 with β thalassemia major, two with transfusion-dependent thalassemia intermedia, six with Eβ thalassemi, two with α thalassemia) except three patients with sickle cell disease. Patients were moderately iron loaded with serum ferritin values of 3243 ± 2280 ng/ml and LIC values of 13.6 ± 7.4 mg/g dry weight. Nineteen patients were positive for hepatitis B or C. MRI assessment of liver R2 and R2* were performed at baseline, 12 weeks, 24 weeks, 48 weeks, 72 weeks, and 96 weeks. All data were collected on 1.5 Tesla scanners using phase array torso or cardiac coils. Liver R2 data were collected and processed using the Ferriscan® protocol and analyzed by a central core laboratory (Resonance Health, Western Australia). Briefly, Ferriscan acquisitions use a matrix of 192 × 256 points, slice thickness 10 mm, field of view 360 × 270 mm, repetition time 1000 ms, and echo times of 6, 9, 12, 15, and 18 ms. R2 values were converted to LIC values using a previously published [12] and independently corroborated [13] calibration curve.

Liver R2* data from different sites were assessed according to local standard of care. All of the centers, except one, used a multiple echo gradient echo sequence with 8–16 echoes and a minimum echo time less than or equal to 1.2 ms. One center (7 subjects) used a multiple breath-hold single echo technique with a minimum echo time of 1 ms. All images were processed centrally at Children's Hospital Los Angeles. Since patients with LIC > 30 mg/g were excluded from the clinical trial, none of the studies had to be discarded for inadequate signal. Signal decay was fit to an exponential term plus a constant and converted to an LIC using a published [13] and independently corroborated [14] calibration curve. Intraobserver and interobserver variability for our processing tools have been previously reported at 0.75% and 2.5%, respectively [17].

ICE may be calculated using the following equation:

$$\text{ICE} = (\text{TII}-\Delta\text{TBI})/(\text{dose x binding capacity}) \quad (1)$$

where TII is the total iron received by transfusions, ΔTBI is the change in total body iron balance, dose represents the interval weight of drug consumed, and binding capacity reflects the grams of iron removed per gram of drug. The binding capacity is simply the ratio of the molecular weight of iron (55.8 g/mole) divided by the product of the molecular weight of deferitazole (400 g/mole) and the number of chelator molecules required to remove one iron molecule (two).

LIC was converted into a change in total body iron (TBI) concentration using the relationship reported by Angelluci et al. [4].

$$\text{TBI} = \text{LIC} \times 10.6\ \text{g/kg} \quad (2)$$

where TBI is reported in mg of iron per kilogram of patient weight and LIC has units of mg/g dry weight of liver.

Using this relationship, we can recast [1] into the following:

$$\text{ICE} = (\text{TII}_\text{D} \times \Delta t-\Delta\text{LIC} \times 10.6\ \text{g/kg} \times \text{Wt})/(\Delta t \times \text{dose}_\text{D} \times \ \text{Wt}) \times (\text{binding capacity}) \quad (3)$$

where TII_D is the TII expressed in mg/day, Δt is the elapsed time in days, Wt is the body weight in kilograms, and dose_D is the actual amount of drug consumed per kilogram per day. Time and weight were explicitly introduced into Eq. (3) to calculate efficiency over multiple observation intervals, allowing corrections for changes in dose, transfusion intensity, and body weight.

The transfusion volume and hematocrit for each unit received by the patient was documented, and TII was calculated as the product of the transfusion volume, transfusion hematocrit and a scaling factor of 1.08 [18]. Drug consumption was meticulously monitored by monthly pill count. Weight was recorded at monthly study visits. Using these data, and the LIC's measured by R2 and R2*, we calculated ICE estimates between the intervals of 0–12, 12–24, 0–24, 24–48, 0–48, and 48–96 weeks.

Thus, by Eq. (3), accurate estimates of LIC should produce consistent and physiologically reasonable ICE estimates. Since a drug cannot bind more iron than its binding sites, nor bind less iron than none, ICE estimates should range from zero to one, unless corrupted by large LIC measurement errors. Furthermore, ICE estimates should not vary dramatically over time, as long as the TII and drug dose are relatively stable. Thus variance in ICE over time was used as a surrogate for LIC measurement robustness. Bland–Altman analysis was performed between ICE estimates calculated at 12 week (0–12 week versus 12–24 week), 24 week (0–24 week versus 24–48 week), and 48 week (0–48 week versus 48–96 week) intervals. Two sample variance test was used to compare the limits of agreement between R2 and R2* at each time-point.

To place MRI-derived chelation ICE estimates into proper context, we also performed simulation experiments to predict how well liver biopsy would perform in estimating iron balance. Liver biopsy represents a random variable, whose mean value tracks the total body iron [4], but has an uncertainty introduced because liver iron is not distributed homogeneously (sampling error) and because the procedures for measuring iron from tissue samples are imperfect (assay error). Estimates of the sampling error magnitude in liver biopsy vary considerably [6–8], so we chose to model them over a broad range (with coefficient of variation (CoV) of 0%, 10%, 20%, 30% and 40%). The assay error represents variability across different metal laboratories. Inductively coupled mass spectrometry is inherently extremely precise (CoV < 5%), however identical liver samples will yield different values at different centers. A multicenter comparison from 40 sites across Australia estimated a CoV among reference laboratories of 12% [19]. Since the sampling and assay variance are independent, the resultant CoV for LIC estimates by biopsy were 12%, 15.6%, 23.3%, 32.3%, and

41.8% respectively, with assay error becoming insignificant as sampling uncertainty increased.

Fig. 1 demonstrates a flow chart for how we performed these mock experiments. We chose the model patients to have similar LIC, TII, drug consumption, ICE and body weight values as our study population. To derive these values, deferitazole data from baseline to 48 weeks were fit to normal or lognormal distributions where appropriate. We then created 100,000 simulated patients, each having unique initial LIC, ICE, TII_D, Wt, and $dose_D$ representative of the patients in the trial. For a given interval, the true change in LIC was completely predictable from these parameters using Eq. (3). However, liver biopsy estimates of initial and final LIC are imperfect, corrupted with white Gaussian multiplicative noise as follows:

$$LIC_{biopsy} = LIC_{true} * (1 + N(0, \ CoV)) \qquad (4)$$

where N(0, CoV) is drawn from a zero mean Gaussian distribution with a standard deviation equal to the biopsy coefficient of variability. Using the LIC_{biopsy} values, we derived an estimate of ICE_{biopsy} that we compared to true ICE. We also evaluated longitudinal stability of ICE_{biopsy} by simulating paired intervals of 12, 24, and 48 weeks where all simulation parameters were held constant. Pairs were plotted as scattergrams in a similar manner as for ICE_{R2} and ICE_{R2*}. If there were no measurement error, these points would form a line of identity. However, errors in LIC_{biopsy} introduce scatter that we estimated using Bland–Altman statistics. Two-sample variance test was used to compare ICE estimates derived from observed (R2, R2*) and simulated (biopsy) LIC data. Simulation experiments were performed at three study durations (12 weeks, 24 weeks, 48 weeks) and five biopsy sampling errors (0%, 10%, 20%, 30%, 40%).

3. Results

Fig. 2 compares the ICE estimates generated by the actual R2 and R2* measurements collected in the clinical trial. ICE was estimated using LIC_{R2} and LIC_{R2*} measured on three difference time scales, 12 weeks (0–12 versus 12–24), 24 weeks (0–24 versus 24–48), and 48 weeks (0–48 versus 48–96). Overall agreement improves with longer observation periods. This is a natural consequence of Eq. (1) because the uncertainty in TBI introduced by poor LIC estimates is static while transfusional iron intake and drug consumed grow over time. When R2 ICE is compared on a 12 week time-scale, ICE estimates were unbiased but 0–12 and 12–24 week ICE values were uncorrelated, having a standard deviation of 0.44 (Table 1). More importantly, 22% of R2 ICE estimates were negative, indicating that LIC_{R2} measurements predicted a rise in total body iron greater than the entire transfusional iron intake. R2 ICE performed better when calculated at 24 week intervals, producing only 9% nonphysiologic (negative) ICE values and lower standard deviation on Bland–Altman analysis (Table 1), but still no correlation was observed between 0–24 and 24–48 week ICE estimates. By 48 week observation intervals, only one negative efficiency was calculated, and a linear relationship was clearly apparent ($r^2 = 0.23$, $p < 0.0125$).

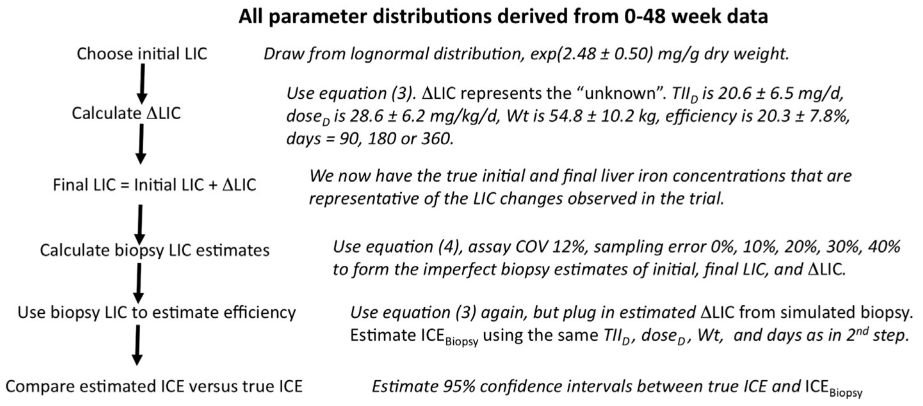
R2* ICE estimates performed better than R2 ICE metrics, particularly on shorter time scales. At twelve week observation intervals, only 5% of measurements were negative; this declined to 2% and 0% at 24 and 48 observation intervals. Linear relationships were significant at 24 and 48 week intervals ($r^2 = 0.27$, $p = 0.0006$ and $r^2 = 0.26$, $p = 0.008$, respectively). On Bland–Altman analysis, R2* ICE measurements exhibited lower variance that R2 ICE at all time points, although the difference did not reach statistical significance at 48 week observation intervals.

Since liver biopsy data were not collected during this clinical trial, it was necessary to simulate liver biopsy data at 0, 12, 24, 48, and 96 weeks as outlined in Fig. 1. Fig. 3 summarizes the simulated ICE metrics at 12 week, 24 week, and 48 week timescales assuming zero liver biopsy sampling error (best case). If LIC estimates had zero measurement error, ICE estimates would populate a straight line connecting the origin to the upper right hand corner. While "perfect" liver biopsy begins to approach this ideal on a 48 week time-scale, many points fall outside the upper right hand quadrant during shorter term observations; this is because the iron assay error of 12% is significant on these shorter time-scales. As observed for the trial R2 and R2* data, ICE estimates for liver biopsy were more robust at 24 and 48 intervals.

Table 1 summarizes the standard deviation for measured (R2, R2*) and simulated (liver biopsy) ICE estimates. At 48 weeks, R2, R2*, and "perfect" liver biopsy exhibited statistically identical variance (indicated by bold type). However, both MRI metrics were superior to any realistic sampling error for liver biopsy (CoV 10%–40%). At shorter measurement intervals, R2* produced the lowest variance ICE estimates. Although statistical separation between R2* and perfect liver biopsy ICE estimates was not attained at 12 weeks, this was driven by two outliers in the R2* ICE estimates. Exclusion of these two points reduced the standard deviation of 12 week R2* efficiency estimates to 15.6%.

Using our simulation paradigm, it was also possible to compare ICE estimates calculated by liver biopsy to true ICE values. By doing so, we could quantitatively assess the impact of observation interval and sampling variability.

Fig. 4 demonstrates the ability of "perfect" liver biopsy to represent true ICE; 95% confidence intervals are displayed for a sampling interval of 12 weeks, 24 weeks, and 48 weeks. At 48 weeks, "perfect" liver biopsy accurately characterizes iron balance in the body, with 95% confidence intervals for ICE estimates



All parameter distributions derived from 0-48 week data

Choose initial LIC — *Draw from lognormal distribution, exp(2.48 ± 0.50) mg/g dry weight.*

Calculate ΔLIC — *Use equation (3). ΔLIC represents the "unknown". TII_D is 20.6 ± 6.5 mg/d, $dose_D$ is 28.6 ± 6.2 mg/kg/d, Wt is 54.8 ± 10.2 kg, efficiency is 20.3 ± 7.8%, days = 90, 180 or 360.*

Final LIC = Initial LIC + ΔLIC — *We now have the true initial and final liver iron concentrations that are representative of the LIC changes observed in the trial.*

Calculate biopsy LIC estimates — *Use equation (4), assay COV 12%, sampling error 0%, 10%, 20%, 30%, 40% to form the imperfect biopsy estimates of initial, final LIC, and ΔLIC.*

Use biopsy LIC to estimate efficiency — *Use equation (3) again, but plug in estimated ΔLIC from simulated biopsy.* Estimate ICE_{Biopsy} using the same *TII_D, $dose_D$, Wt, and days as in 2^{nd} step.*

Compare estimated ICE versus true ICE — *Estimate 95% confidence intervals between true ICE and* ICE_{Biopsy}

Fig. 1. Flow chart profiling the simulation process.

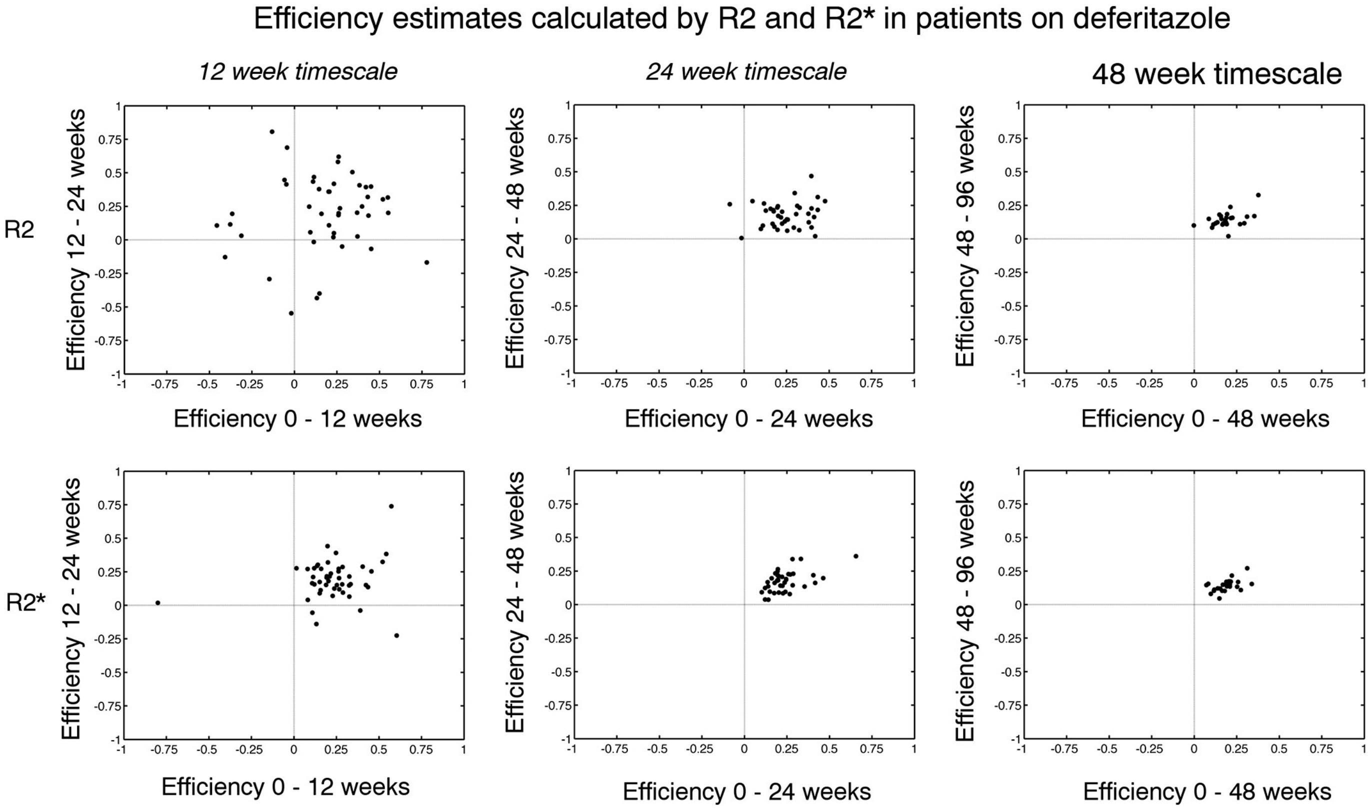


Fig. 2. Scatterplot of estimated ICE across different observation intervals, 12 weeks (left), 24 weeks (middle) and 48 weeks (right), calculated by R2 (top row) and R2* (bottom row). All R2* efficiency estimates at 24 and 48 weeks reside in the upper right hand quadrant.

of ± 7%. However, the ability to predict ICE on a shorter time-scale is markedly compromised, particularly at 12 week intervals. For example, at the average ICE observed in the trial (20.6%), the 95% confidence intervals for 12 week observations extends from an efficiency of −10% to 51%.

Fig. 5 demonstrates biopsy-estimated ICE versus true ICE, calculated over 48 weeks, for realistic sampling errors [6–8]. Even if one uses a fairly conservative sampling error estimate of 20%, the 95% confidence intervals for biopsy-estimated ICE estimates are 6.8% to 35.1%, respectively, for a “true” ICE value of 20.6%. When true ICE is one standard deviation below the mean (10.4%), many of the biopsy-derived ICE estimates would be negative. Thus for any given individual, the ability of physically-achievable liver biopsy to estimate ICE is quite poor even on a one year time-scale.

4. Discussion

No gold standard exists for measuring liver iron concentration. We demonstrate that the longitudinal variance in ICE estimates is a powerful metric to evaluate LIC measurement performance in the absence of a gold standard. This approach is particular important because it has become ethically challenging to perform liver biopsy for iron quantification as MRI relaxometry techniques have improved.

By Eq. (3), LIC measurement errors propagate into the uncertainty of chelator efficiency proportionately to the LIC difference. If LIC measurement errors are consistent over time, they cancel out in the calculation of chelator efficiency. In contrast, independent LIC measurement errors produce additive uncertainty with respect to efficiency estimates. This mirrors the well-known principle that bias is better tolerated in clinical practice than variability.

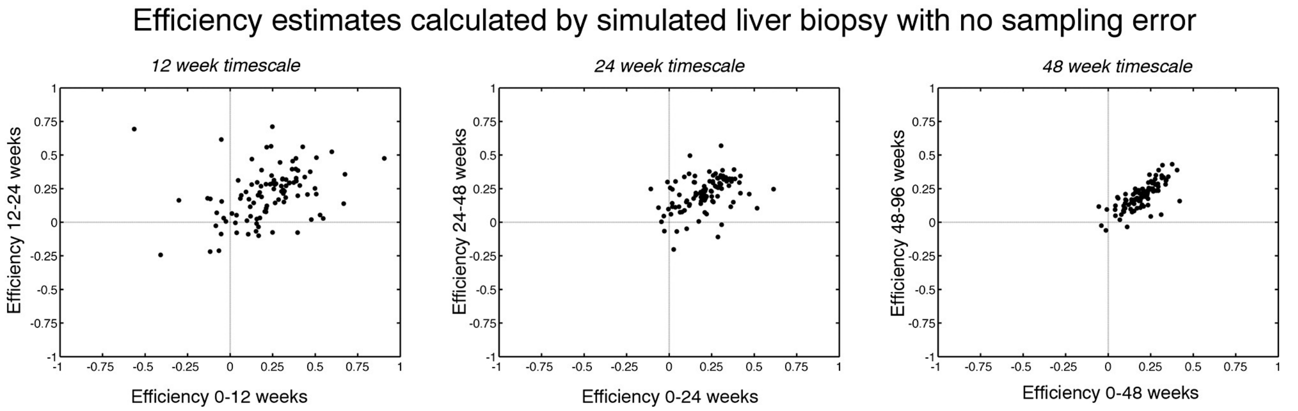
Liver biopsy has no known patient specific biases, but sampling variability estimates between 9% and 44% have been reported [6–8,20,21]. Specimen size, patient age, chelation history, and presence of fibrosis all impact sampling error. Systematic bias and additional variability can be added if the sample is embedded in paraffin because xylene dewaxing procedures remove a variable amount of tissue lipids depending on their intensity [9]. As a result, we chose our model's sampling variability to span all reasonable values. Additional, statistically independent, variance is introduced through the biochemical assay. Although relatively modest, estimated at 12% across a large sample of metal laboratories [19], it is not insignificant when trying to monitor iron chelation therapy within individuals. Inductively coupled mass spectrometry (ICP-MS) may yield slightly lower CoV (7%) within a single metal laboratory [13], however no systemic interlaboratory study has been reported for ICP-MS to date.

In contrast, MRI has exceeding low sampling variability. Slice-to-slice variability of 10% and interstudy variability of 7% has been reported for Ferriscan® measurements [12]. For liver R2*, slice to slice, interobserver, intermachine, and interstudy variability estimates vary slightly among studies, but the cumulative errors are generally 5–7%

Table 1
Standard deviation of ICE across observation intervals.

	12 Weeks	24 Weeks	48 Weeks
R2	44.8%	14.8%	**7.4%**
R2*	**22.9%**	**9.3%**	**5.7%**
Biopsy, sampling error 0%	**25.9%**	12.8%	**6.7%**
Biopsy, sampling error 10%	32.8%	16.5%	8.7%
Biopsy, sampling error 20%	49.1%	24.9%	13.0%
Biopsy, sampling error 30%	68.1%	34.5%	18.0%
Biopsy, sampling error 40%	88.1%	44.6%	23.2%

Bold indicates most robust estimate ($p < 0.05$ by variance test) for each interval.


Efficiency estimates calculated by simulated liver biopsy with no sampling error

12 week timescale

24 week timescale

48 week timescale

Fig. 3. Scatterplot of estimated ICE across different observation intervals, 12 weeks (left), 24 weeks (middle) and 48 weeks (right), calculated by "perfect" liver biopsy. Points lying outside the upper right hand quadrant are common at 12 and 24 weeks and represent physiologically impossible values.

[13,17,22–24]. However all MRI relaxivity metrics exhibit errors related to patient-specific differences in tissue iron distribution and magnetic susceptibility [25]. As a result, MRI is not inherently more "accurate" than liver biopsy for predicting true LIC on any single study until biopsy sampling variability is around 20% or higher.

During serial studies, though, the patient specific bias observed in R2 and R2* LIC measurements is minimized (cancel one another during trending) while the random errors by biopsy are amplified. MRI's superb interstudy reproducibility trumps its calibration errors for prediction of individual response to therapy. Clinically, therapeutic response (i.e. the direction and magnitude of change over time) is more important than absolute iron levels, contributing to the rising popularity of MRI in monitoring iron chelation therapy [11,26]. All of these points are in addition to patient acceptance of the method, and the ability to make measurements as frequently as every six months.

R2 and R2* provided equally robust ICE estimates at 48 week intervals, concordant with the excellent cross-sectional correlation observed between these techniques and liver biopsy [12–14,27]. However, R2* measurements had superior robustness during short-term observations. The reason for this disparity is that R2 measurements are exquisitely sensitive to iron deposits having a size of young siderosomes [25,28]. As lysosomes grow or aggregate together, the radiofrequency pulse produces static refocusing, thereby lowering R2 relaxation [25]. Increasing length scale of iron deposits is what yields the downward curvature in the R2-iron calibration curve [25]. R2* measurements lack static refocusing that occurs at larger iron scales [29], causing R2* values to be larger than R2 but retain linearity with iron burden [30,31]. As a result, R2* measurements are much less sensitive than R2 measurements to acute changes in tissue iron distribution [28,30,32] or proton mobility [25,32], responding primarily to total tissue magnetic susceptibility. In this trial, all patients were placed on a novel agent, and some were also under-dosed for the first 24 weeks, producing disequilibrium in the length scale of tissue iron [32]. While 12 week observation intervals are not typically performed in clinical practice, rapid dose adjustment may be desirable in high-risk situations or in dose-finding clinical trials. R2* was also superior for examinations performed at 24 week intervals, but the difference was smaller; R2 assessment was still at least as accurate as simulated liver biopsy results at 24 week intervals. Semiannual (24 weekly) MRI assessments are clinically relevant, being used in high-risk patients as well as clinical trials of iron chelation therapy.

In fact, we propose that the standard deviation of ICE estimates on three to six month time-scales may represent the gold standard to

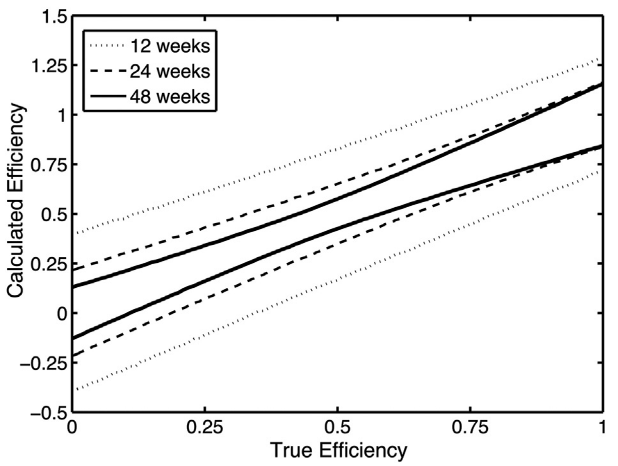


Fig. 4. Errors in ICE estimate calculated using "perfect" liver biopsy (no sampling error) when calculated at different observation intervals. Lines represent 95% confidence intervals for different observations lengths (12, 24, and 48 weeks). Agreement between predicted and true ICE at 12 weeks is too poor to be clinically useful. ICE is unitless but can be converted to percent by multiplying by 100.

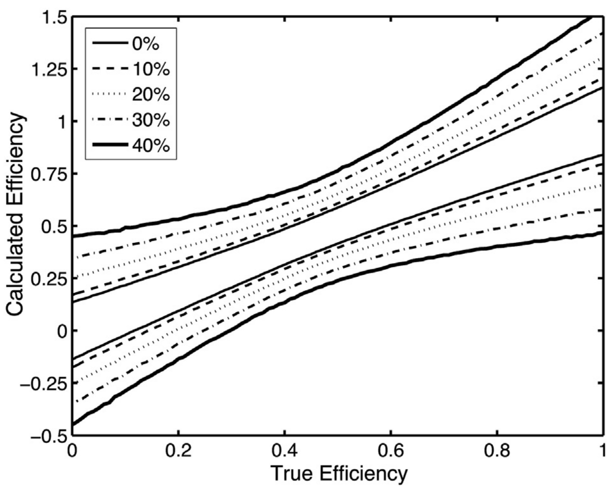


Fig. 5. Errors in ICE estimates calculated using "imperfect" liver biopsy. Lines represent 95% confidence intervals for sampling CoV ranging from 0% to 40%; agreement worsens proportionally to sampling error. Observation interval is 48 weeks for all comparisons.

Please cite this article as: Wood JC, et al, Liver MRI is more precise than liver biopsy for assessing total body iron balance: a comparison of MRI relaxometry with simulated liver biopsy results, Magn Reson Imaging (2015), http://dx.doi.org/10.1016/j.mri.2015.02.016

evaluate any LIC metric. Shorter intervals may overemphasize transient changes in iron distribution, while longer intervals de-emphasize contributions of LIC measurement errors compared with transfusional iron burden and drug consumption.

This study had several important limitations. Firstly, not all of the patients completed all of the study time-points, so we cannot exclude a potential sampling bias. Secondly, our assumption of constant ICE over time is imperfect. While chelator dose was constant for the first 24 weeks, dose escalations occurred across the rest of the trial. Although we could not identify a systematic relationship between dose and ICE, we cannot exclude a contribution to the observed variability in these measurements. Thirdly, we cannot control for patient specific differences in iron absorption or excretion. Fourthly, the Angelucci relationship does not control for any changes in the liver volume, nor has it been validated for diseases other than thalassemia. Fortunately, all of the aforementioned uncertainties will affect LIC R2 and LIC R2* equally and cause our simulations to overestimate, rather than underestimate, liver biopsy performance. Thus while we could not actually perform liver biopsy in these patients, because of ethical constraints, our simulations represent truly “best case” performance of liver biopsy. Even so, we can conclude the MRI performs as well as physiologically unachievable (zero sampling error) liver biopsy in estimating iron balance in patients. MRI is unconditionally superior to biopsy when sampling variability is equal to values reported in the literature. The intercenter assay error of 12% was reported in 1980 and could potentially be lower with more modern instrumentation. It was also calculated from non-iron loaded liver and could be different for clinical specimens. If a lower assay error were used, one would expect better performance of biopsy for CoV < 20%. However, assay error becomes negligible with respect to sampling errors larger than 20%, such as would be expected in this adult population with high prevalence of hepatitis.

In summary, we demonstrate that MRI relaxometry produces LIC estimates that are better for monitoring individual patient response to iron chelation therapy than values produced by physically-achievable liver biopsy. We conclude that well-controlled MRI relaxometry metrics should replace liver biopsy as a surrogate for chelator effectiveness in clinical trials and that indications for liver biopsy should be restricted to assessment of tissue histology.

Conflict of interest statement

Dr. Wood serves as a consultant to Shire, ApoPharma, and Biomed Informatics. He has research funding from Shire and the National Institutes of Health. Dr. Wood has received speaker reimbursements from Novartis and ApoPharma. Hugh Rienhoff was the CEO of Ferrokin Biosciences which is now a wholly owned subsidiary of Shire. Pingao Zhang and Walid Abi Saab are employees of Shire. Dr. Ellis Neufeld received research funding from Shire and serves as a consultant.

Acknowledgements

This work was supported by Shire Pharmaceuticals and the National Institute of Health (1R01DK097115-01A1). Data from this study were presented as a poster at the American Society of Hematology Meetings, New Orleans, Dec 7–10, 2013.

References

[1] Bird RJ, Kenealy M, Forsyth C, Wellwood J, Leahy MF, Seymour JF, et al. When should iron chelation therapy be considered in patients with myelodysplasia and other bone marrow failure syndromes with iron overload? Intern Med J 2012;42(4):450–5.

[2] Porter J, Garbowski M. Consequences and management of iron overload in sickle cell disease. Hematology Am Soc Hematol Educ Program 2013;2013:447–56.

[3] Brittenham GM, Griffith PM, Nienhuis AW, McLaren CE, Young NS, Tucker EE, et al. Efficacy of deferoxamine in preventing complications of iron overload in patients with thalassemia major. N Engl J Med 1994;331(9):567–73.

[4] Angelucci E, Brittenham GM, McLaren CE, Ripalti M, Baronciani D, Giardini C, et al. Hepatic iron concentration and total body iron stores in thalassemia major. N Engl J Med 2000;343(5):327–31.

[5] Olivieri NF, Brittenham GM. Iron-chelating therapy and the treatment of thalassemia. Blood 1997;89(3):739–61.

[6] Ambu R, Crisponi G, Sciot R, Van Eyken P, Parodo G, Iannelli S, et al. Uneven hepatic iron and phosphorus distribution in beta-thalassemia. J Hepatol 1995;23(5):544–9.

[7] Emond MJ, Bronner MP, Carlson TH, Lin M, Labbe RF, Kowdley KV. Quantitative study of the variability of hepatic iron concentrations. Clin Chem 1999;45(3):340–6.

[8] Villeneuve JP, Bilodeau M, Lepage R, Cote J, Lefebvre M. Variability in hepatic iron concentration measurement from needle-biopsy specimens. J Hepatol 1996;25(2):172–7.

[9] Butensky E, Fischer R, Hudes M, Schumacher L, Williams R, Moyer TP, et al. Variability in hepatic iron concentration in percutaneous needle biopsy specimens from patients with transfusional hemosiderosis. Am J Clin Pathol 2005;123(1):146–52.

[10] Angelucci E, Baronciani D, Lucarelli G, Baldassarri M, Galimberti M, Giardini C, et al. Needle liver biopsy in thalassaemia: analyses of diagnostic accuracy and safety in 1184 consecutive biopsies. Br J Haematol 1995;89(4):757–61.

[11] Kwiatkowski JL, Kim HY, Thompson AA, Quinn CT, Mueller BU, Odame I, et al. Chelation use and iron burden in North American and British thalassemia patients: a report from the Thalassemia Longitudinal Cohort. Blood 2012;119(12):2746–53.

[12] St Pierre TG, Clark PR, Chua-anusorn W, Fleming AJ, Jeffrey GP, Olynyk JK, et al. Noninvasive measurement and imaging of liver iron concentrations using proton magnetic resonance. Blood 2005;105(2):855–61.

[13] Wood JC, Enriquez C, Ghugre N, Tyzka JM, Carson S, Nelson MD, et al. MRI R2 and R2* mapping accurately estimates hepatic iron concentration in transfusion-dependent thalassemia and sickle cell disease patients. Blood 2005;106(4):1460–5.

[14] Hankins JS, McCarville MB, Loeffler RB, Smeltzer MP, Onciu M, Hoffer FA, et al. R2* magnetic resonance imaging of the liver in patients with iron overload. Blood 2009;113(20):4853–5.

[15] Bergeron RJ, Streiff RR, Wiegand J, Luchetta G, Creary EA, Peter HH. A comparison of the iron-clearing properties of 1,2-dimethyl-3-hydroxypyrid-4-one, 1,2-diethyl-3-hydroxypyrid-4-one, and deferoxamine. Blood 1992;79(7):1882–90.

[16] Neufeld EJ, Galanello R, Viprakasit V, Aydinok Y, Piga A, Harmatz P, et al. A phase 2 study of the safety, tolerability and pharmacodynamics of FBS0701, a novel oral iron chelator, in transfusional iron overload. Blood 2012;199(14):3263–8.

[17] Saiviroonporn P, Viprakasit V, Sanpakit K, Wood JC, Krittayaphong R. Intersite validations of the pixel-wise method for liver R2* analysis in transfusion-dependent thalassemia patients: a more accessible and affordable diagnostic technology. Hematol Oncol Stem Cell Ther 2012;5(2):91–5.

[18] Cohen AR, Glimm E, Porter JB. Effect of transfusional iron intake on response to chelation therapy in beta-thalassemia major. Blood 2008;111(2):583–7.

[19] Koh TS, Benson TH, Judson GJ. Trace element analysis of bovine liver: interlaboratory survey in Australia and New Zealand. J Assoc Off Anal Chem 1980;63(4):809–13.

[20] Barry M, Sherlock S. Measurement of liver-iron concentration in needle-biopsy specimens. Lancet 1971;1(7690):100–3.

[21] Kreeftenberg HG, Koopman BJ, Huizenga JR, van Vilsteren T, Wolthers BG, Gips CH. Measurement of iron in liver biopsies–a comparison of three analytical methods. Clin Chim Acta 1984;144(2–3):255–62.

[22] Westwood MA, Anderson LJ, Firmin DN, Gatehouse PD, Lorenz CH, Wonke B, et al. Interscanner reproducibility of cardiovascular magnetic resonance T2* measurements of tissue iron in thalassemia. J Magn Reson Imaging 2003;18(5):616–20.

[23] Westwood MA, Firmin DN, Gildo M, Renzo G, Stathis G, Markissia K, et al. Intercentre reproducibility of magnetic resonance T2* measurements of myocardial iron in thalassaemia. Int J Cardiovasc Imaging 2005;21(5):531–8.

[24] Kirk P, He T, Anderson LJ, Roughton M, Tanner MA, Lam WW, et al. International reproducibility of single breathhold T2* MR for cardiac and liver iron assessment among five thalassemia centers. J Magn Reson Imaging 2010;32(2):315–9.

[25] Ghugre NR, Wood JC. Relaxivity-iron calibration in hepatic iron overload: probing underlying biophysical mechanisms using a Monte Carlo model. Magn Reson Med 2011;65(3):837–47.

[26] Modell B, Khan M, Darlison M, Westwood MA, Ingram D, Pennell DJ. Improved survival of thalassaemia major in the UK and relation to T2* cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2008;10(1):42–50.

[27] Garbowski MW, Carpenter JP, Smith G, Roughton M, Alam MH, He T, et al. Biopsy-based calibration of T2* magnetic resonance for estimation of liver iron concentration and comparison with R2 Ferriscan. J Cardiovasc Magn Reson 2014;16:40–51.

[28] Wood JC, Fassler J, Meade T. Mimicking liver iron overload using liposomal ferritin preparations. Mag Res Med 2004;51(3):607–11.

[29] Weisskoff RM, Zuo CS, Boxerman JL, Rosen BR. Microscopic susceptibility variation and transverse relaxation: theory and experiment. Magn Reson Med 1994;31(6):601–10.

[30] Tanimoto A, Oshio K, Suematsu M, Pouliquen D, Stark DD. Relaxation effects of clustered particles. J Magn Reson Imaging 2001;14(1):72–7.

ARTICLE IN PRESS

[31] Tanimoto A, Pouliquen D, Kreft BP, Stark DD. Effects of spatial distribution on proton relaxation enhancement by particulate iron oxide. J Magn Reson Imaging 1994;4(5):653–7.

[32] Wood J, Aguilar M, Otto-Duessel M, Nick H, Nelson M, Moats R. Influence of iron chelation therapy on R1 and R2 calibration curves in gerbil liver and heart. Mag Res Med 2008;60(1):82–9.

Please cite this article as: Wood JC, et al, Liver MRI is more precise than liver biopsy for assessing total body iron balance: a comparison of MRI relaxometry with simulated liver biopsy results, Magn Reson Imaging (2015), http://dx.doi.org/10.1016/j.mri.2015.02.016