Received: 11 October 2023 | Revised: 20 February 2024 | Accepted: 16 March 2024

DOI: 10.1002/mrm.30100

RESEARCH ARTICLE

Magnetic Resonance in Medicine

The "hidden noise" problem in MR image reconstruction

Jiayang Wang | **Di An** | **Justin P. Haldar**

Signal and Image Processing Institute, Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California, USA

Correspondence
Jiayang Wang, University of Southern California, University Park Campus, 3710 McClintock Ave, EEB 400, Los Angeles, CA 90089, USA.
Email: jiayangw@usc.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01-MH116173, R01-NS074980, R56-EB034349; Ming Hsieh Institute for Research on Engineering-Medicine for Cancer; University of Southern California (Annenberg Graduate Fellowship)

Abstract

Purpose: The performance of modern image reconstruction methods is commonly judged using quantitative error metrics like root mean squared-error and the structural similarity index, which are calculated by comparing reconstructed images against fully sampled reference data. In practice, the reference data will contain noise and is not a true gold standard. In this work, we demonstrate that the "hidden noise" present in reference data can substantially confound standard approaches for ranking different image reconstruction results.

Methods: Using both experimental and simulated k-space data and several different image reconstruction techniques, we examined whether there was correlation between performance metrics obtained with typical noisy reference data versus those obtained with higher-quality reference data.

Results: For conventional performance metrics, the reconstructions that matched best with the higher-quality reference data were substantially different from the reconstructions that matched best with typical noisy reference data. This leads to suboptimal reconstruction results if the performance with respect to noisy reference data is used to select which reconstruction methods/parameters to employ. These issues were reduced when employing alternative error metrics that better account for noise.

Conclusion: Reference data containing hidden noise can substantially mislead the ranking of image reconstruction methods when using conventional error metrics, but this issue can be mitigated with alternative error metrics.

KEYWORDS
image quality assessment, image reconstruction, noise

1 | INTRODUCTION

In recent years, it has become increasingly common to judge the quality of MRI reconstruction methods using quantitative error metrics like root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM).[1] Driven in part by trends in machine learning,[2–4] it has also become increasingly common that these quantitative error metrics are used for the empirical tuning of image reconstruction parameters and/or to provide guidance on which image reconstruction methods should be employed in practical applications. (Interestingly, the use of these kinds of quantitative error metrics was not popular in the early MRI reconstruction literature—interested readers are referred to the bibliography of Reference 5 for historical context).

RMSE, MAE, and SSIM are all examples of "full-reference" error metrics,[1] and are intended to be

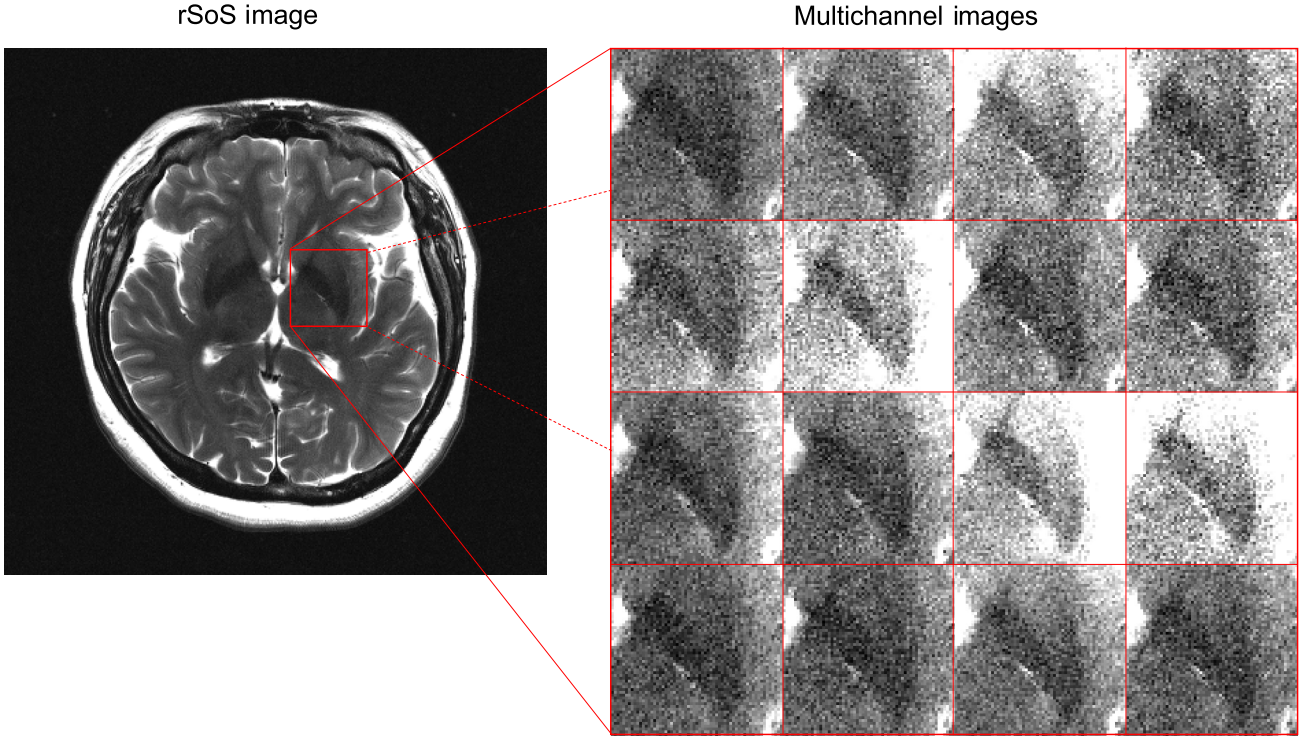

FIGURE 1 A representative multichannel T2-weighted brain image from the fastMRI dataset. Left: root sum-of-squares (rSoS) image. Right: Depiction of the 16 original individual channels that were combined to form the rSoS image. It is visually obvious that the individual channels each contain substantial amounts of noise.

used to evaluate the degree of similarity between a reconstructed image and a "gold standard" reference image. However, it is never possible to experimentally measure a true gold standard in MRI, because real data will always contain thermal noise. As such, it has become standard procedure to calculate measures like RMSE, MAE, and SSIM using an image formed from noisy Nyquist-sampled k-space data as a reference, often based on the implicit assumption that the noise should be small and have a negligible effect on the calculated values. We refer to the noise in a noisy reference image as "hidden noise," since conventional error metrics will treat it as a desired part of the reference image rather than properly treating it as undesirable error.

Importantly, hidden noise can be substantial, even when it is not visually obvious. To illustrate, Figure 1 shows a typical multichannel brain MRI dataset from the widely used fastMRI database.[7] While the root sum-of-squares (rSoS) image (a common choice of reference image in the modern literature) visually appears to have excellent signal-to-noise ratio (SNR), closer examination demonstrates that there is actually substantial noise present in the individual channels. This implies that the pristine appearance of the rSoS image is misleading, and hides a substantial noise bias. Note also that rSoS images should follow the non-central chi (NCC) distribution,[8–11] with an expected value that is biased away from the value that would be observed with noiseless data. Hidden noise is also evident in k-space, as illustrated in Figure 2, which demonstrates that the SNR is small (<2) for substantial portions of k-space and suggests that high-frequency information may be unreliable in the original raw data.

In this work, we investigate how hidden noise can impact the assessment of image reconstruction performance when typical noisy Nyquist-sampled rSoS images are used as reference images. Importantly, we observe that the effects of hidden noise can cause incorrect ranking of image reconstruction results, leading to suboptimal reconstruction quality when standard performance metrics are used to optimize reconstruction performance. However, we also demonstrate that these confounds can be reduced when employing alternative error metrics that better account for noise.

A preliminary account of portions of this work was previously presented at a recent conference.[14]

2 | THEORY

Traditional full-reference image quality measures like RMSE, MAE, and SSIM were designed for scenarios where the reference image is pristine, and does not possess the bias or variability characteristics of a noisy reference image. When using a noisy reference image in

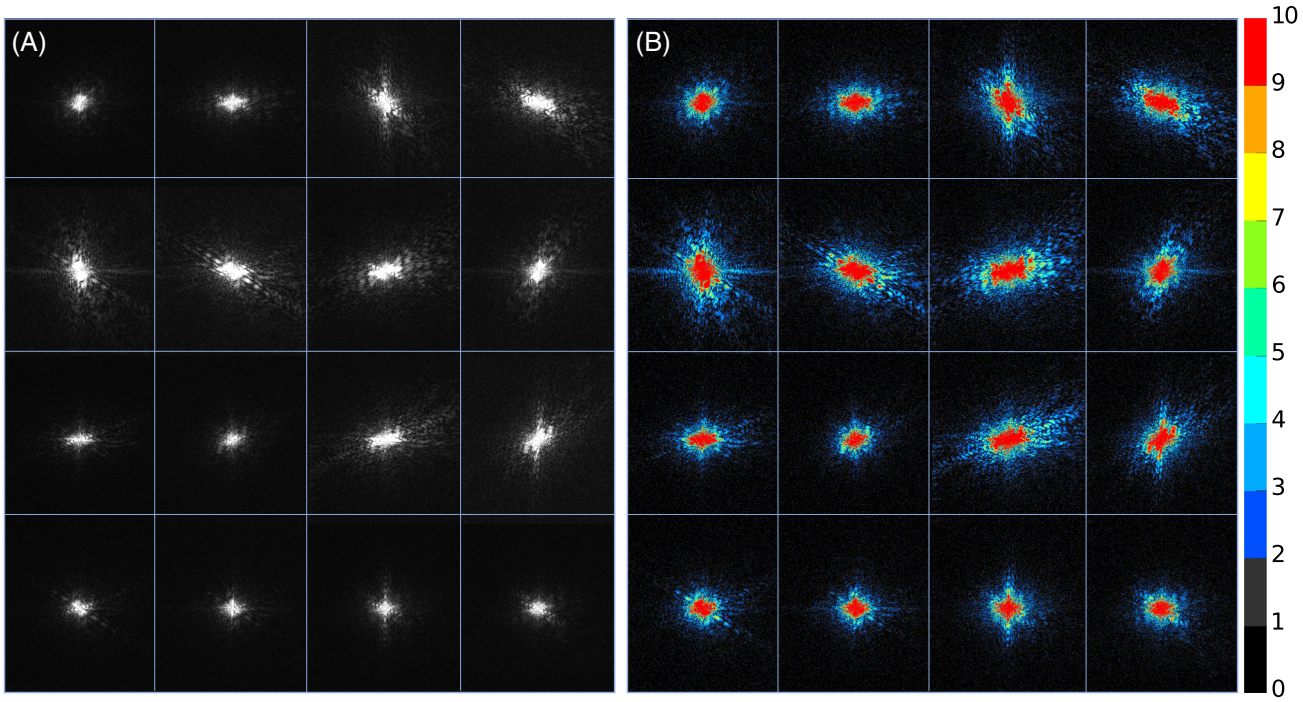

FIGURE 2 (A) 16-channel k-space data corresponding to the same dataset from Figure 1. (B) The same data from (A), but normalized by the noise variance of each channel and color-coded to more easily visualize the signal-to-noise ratio (SNR) characteristics in different regions of k-space. In all channels, there are substantial regions of k-space with SNR<2.

place of a pristine reference image, treating the noise as part of the desired gold standard may lead to RMSE, MAE, and SSIM values that are easily misinterpreted, and properly accounting for noise in the reference image may be essential for an accurate assessment of image quality.

Fortunately, methods to account for noise bias and variability are well explored in fields like statistical signal processing and estimation theory.[15] In these fields, consistency with noisy data would not typically be measured with metrics like RMSE, MAE, or SSIM except in circumstances where these metrics happened to coincide with characteristics of the noise distribution. Instead, standard practice would be to measure data consistency using statistically informed ("noise aware") metrics like the negative log-likelihood function. If $\mathbf{x}$ represents an estimate of the unknown parameters (e.g., an MRI image reconstruction), then the negative log-likelihood function assuming noisy reference measurements $\mathbf{y}$ would be given by $-\ln p(\mathbf{y}|\mathbf{x})$, where $p(\mathbf{y}|\mathbf{x})$ represents the conditional probability (depending directly on noise modeling assumptions) of measuring the noisy data $\mathbf{y}$ under the assumption that the estimated values $\mathbf{x}$ were perfectly accurate. The negative log-likelihood function is one of the key elements of the statistical signal processing and estimation theory toolbox, and for example is the sole function used to measure consistency with a noisy measurement in popular methods like maximum-likelihood estimation, penalized maximum-likelihood estimation, and (Bayesian) maximum a posteriori estimation.[15]

In the following subsections, we review the noise model and the classical negative log-likelihood function for rSoS MRI images, which we use to define a statistically-motivated image quality metric that better accounts for noise.

2.1 | Noise modeling for rSoS MRI images

Without loss of generality, we will describe noise modeling for a simple two-dimensional multichannel MR imaging scenario where data is collected from L distinct channels, although the same ideas generalize naturally to other contexts. For each channel, we assume that noisy k-space samples have been acquired at the Nyquist rate on a fully sampled Cartesian grid of size $N_x \times N_y$, and that a basic inverse discrete Fourier transform (without apodization, zero-paddding, or other forms of processing that could influence the noise distribution) has been applied to the k-space data to produce a noisy image for the ℓth channel $f_\ell[m,n]$ defined on an $N_x \times N_y$ voxel grid. The noisy image $f_\ell[m,n]$ can be decomposed as $f_\ell[m,n] = g_\ell[m,n] + z_\ell[m,n]$, where $g_\ell[m,n]$ represents the ideal image that would have been acquired in the absence of noise, while $z_\ell[m,n]$ represents the noise contribution.

Using standard thermal noise assumptions, the acquisition and reconstruction procedures described above imply that the noise samples at distinct voxel locations will follow independent and identical zero-mean circularly symmetric complex-valued Gaussian distributions.[16]

A noiseless ("gold standard") rSoS reference image is not possible to obtain in practice, but would ideally be given by

$$r_{\text{ideal}}[m,n] = \sqrt{\sum_{\ell=1}^{L} |g_\ell[m,n]|^2}. \tag{1}$$

The practical ("noisy") rSoS reference image that is popular in the modern literature is given by

$$\begin{aligned} r_{\text{noisy}}[m,n] &= \sqrt{\sum_{\ell=1}^{L} |f_\ell[m,n]|^2} \\ &= \sqrt{\sum_{\ell=1}^{L} |g_\ell[m,n] + z_\ell[m,n]|^2}. \end{aligned} \tag{2}$$

If the interchannel noise covariance matrix has been prewhitened such that the complex-valued noise in each channel is uncorrelated with common variance σ^2,[17] then it is straightforward to show that $r_{\text{noisy}}[m,n]$ follows a NCC distribution, and that

$$E\left\{(r_{\text{noisy}}[m,n])^2\right\} = (r_{\text{ideal}}[m,n])^2 + \sigma^2 L, \tag{3}$$

where we have chosen to show the expectation of the square of $r_{\text{noisy}}[m,n]$ because it leads to a much simpler mathematical expression than the exception of $r_{\text{noisy}}[m,n]$ itself (see, e.g., Reference 8 for a complicated expression for the expectation of $r_{\text{noisy}}[m,n]$). It is easy to observe that the values of $r_{\text{noisy}}[m,n]$ will be biased away from the values of $r_{\text{ideal}}[m,n]$, with the amount of bias dependent on the noiseless signal intensity.[8–11] In addition, it can be shown that $r_{\text{ideal}}[m,n]$ will have a signal-dependent (spatially varying) variance,[8–11] meaning that some voxel values will be more or less reliable than others.

Assuming that $r_{\text{recon}}[m,n]$ denotes a reconstructed MRI magnitude image, and that we wish to compare this reconstruction against a noisy rSoS reference image $r_{\text{noisy}}[m,n]$ that obeys the the NCC noise assumptions described above, the corresponding negative log-likelihood would take the form (neglecting unimportant additive constants)[8–11]

$$-\ln p(\mathbf{y}|\mathbf{x}) = \sum_{m,n} \left\{ \begin{array}{l} \frac{1}{\sigma^2}(r_{\text{recon}}[m,n])^2 + (L-1)\ln(r_{\text{recon}}[m,n]) \\ \quad - \ln I_{L-1}\left(\frac{(r_{\text{recon}}[m,n])(r_{\text{noisy}}[m,n])}{0.5\sigma^2}\right) \end{array} \right\}, \tag{4}$$

where $I_L(\cdot)$ is the Lth-order-modified Bessel function of the first kind.

Although the previous description is specific to prewhitened data, it is common within the modern machine learning literature form reference images by applying rSoS directly to images from the original channels without any prewhitening. In this case, the complex-valued noise images $z_\ell[m,n]$ will possess an interchannel noise covariance matrix $\mathbf{\Psi} \in \mathbb{C}^{L\times L}$,[17,18] and the rSoS coil combined image will no longer exactly match the NCC distribution.[10] However, it has been previously demonstrated that the NCC distribution can still be used to closely approximate the distribution that arises from non-prewhitened data[10] as long as the negative log-likelihood function from Equation (4) is evaluated using appropriately-chosen "effective" NCC parameters $\tilde{L}$ and $\tilde{\sigma}$ that differ from the physical L and σ parameters.

2.2 | Noncentral chi error metric

As noted previously, the negative log-likelihood provides a natural "noise aware" metric of data consistency. In this work, we define the noncentral chi error (NCE) metric as a normalized version of the classical negative log-likelihood. Specifically, the NCE metric measures the distance between a reconstruction $\mathbf{x}$ and a noisy reference image $\mathbf{y}$ using

$$\text{NCE}(\mathbf{x},\mathbf{y}) = -\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{y}), \tag{5}$$

where the negative log-likelihoods are calculated using Equation (4), and the second term in Equation (5) is a normalization constant that ensures that $\text{NCE}(\mathbf{y},\mathbf{y}) = 0$. Similar to other commonly used normalization approaches, our choice of normalization for NCE is somewhat arbitrary and not strictly necessary, though helps make the NCE$(\cdot)$ values more consistent and potentially easier to interpret and compare across different datasets. Note that the NCE$(\cdot)$ function can take on both positive and negative real values, with smaller values (i.e., larger negative values) indicating better consistency between the reconstructed image and the noisy reference image.

Evaluation of the NCE metric requires calculation of negative log-likelihood values, which, in the context of nonwhitened data, requires knowledge of $\tilde{\sigma}$ and $\tilde{L}$. Ideally, these parameters could be derived from knowledge of the subject-specific interchannel noise covariance matrix $\mathbf{\Psi}$, which is typically measured by default on most MRI scanners using noise-only calibration scans. However, these noise calibration scans are frequently unavailable from large public repositories of MRI k-space data. For the results shown in this work (and similar to Reference 11),

we find image-dependent maximum likelihood estimates for the values of $\tilde{\sigma}$ and $\tilde{L}$ based on data from background ("noise only") regions of each noisy image.

3 | METHODS

The fact that noise is always present in real MRI data represents a fundamental obstacle to the evaluation of different image quality metrics, and means that practical compromises must be made in order to define "clean" reference images. We have approached this issue in two different ways, one based on averaging and the other based on denoising. Neither of these approaches is perfect, but the two approaches have different limitations and should provide complementary insights.

3.1 | Testing regularized reconstruction using averaged data

Our first set of experiments is based on a slice from a standard fully-sampled in vivo brain MPRAGE dataset acquired at our institution with a 3T Siemens Prisma scanner (roughly 1 mm isotropic resolution over a 210 mm × 154 mm field-of-view, using a 32-channel receiver array). In order to obtain both a typical reference dataset (to emulate standard performance assessment with a noisy reference) and a higher-SNR reference dataset (to more accurately assess true performance), we acquired this data with five averages. A single average was used to create the typical ("noisy") reference, while all averages (with additional singular value decomposition-based denoising, achieved by zeroing the noise-dominated principal components along the channel dimension[19]) were used for the higher-SNR ("clean") reference. While the clean reference image is not perfectly noiseless, it is nevertheless substantially better than the noisy reference.

This MPRAGE data were used to emulate a typical parameter-tuning application in regularized image reconstruction from accelerated k-space data. Specifically, the (noisy) single-average k-space data was retrospectively undersampled, using a uniform one-dimensional undersampling pattern that kept every third line of k-space while also fully sampling the central 16 k-space lines (total acceleration factor = 2.6×). This undersampled data was reconstructed using three different regularized reconstruction methods: SENSE-based parallel imaging[17] with a total variation regularization constraint (SENSE-TV)[20,21]; Autocalibrated LORAKS (AC-LORAKS) reconstruction[22] which imposes phase constraints, support constraints, and parallel imaging constraints (i.e., the LORAKS **S**-matrix[22–24]) using a fixed low-dimensional subspace that has been pre-estimated from the fully sampled center of k-space; and P-LORAKS[22] which imposes the same constraints as AC-LORAKS but updates the estimate of the low-dimensional subspace at each iteration of the optimization algorithm. All three reconstructions were formulated using penalized maximum likelihood principles (i.e., the noise covariance was included in the data consistency term in each case).[17] SENSE-TV was implemented using sensitivity maps estimated with the PISCO algorithm[25] and optimization was performed using the version of Nesterov's algorithm described in References 26 and 27, while AC-LORAKS was implemented using a multiplicative half-quadratic algorithm without FFTs[28] and P-LORAKS was implemented using an additive half-quadratic algorithm.[28] To emulate typical parameter tuning applications, reconstructions were performed using a variety of different reconstruction parameters (i.e., different regularization parameters for SENSE-TV, and different regularization parameters and matrix ranks for AC-LORAKS and P-LORAKS). Reconstructions obtained with each method were coil-combined using rSoS (in the case of SENSE-TV, we used the sensitivity maps to generate multichannel images prior to rSoS combination), and were compared against the "clean" and "noisy" rSoS reference images using conventional error metrics (RMSE, MAE, SSIM) and our new NCE metric.

3.2 | Testing machine learning reconstruction with denoised data

Our second set of experiments is based on 16-channel T2-weighted brain images from the fastMRI database.[7] We selected the central eight brain slices from 425 subjects in the database (3400 slices total), and applied standard filtering and downsampling operations to finish the remaining steps of oversampled analog-to-digital conversion[29] (the data was acquired with 2× oversampling along the readout dimension but was provided in an intermediate raw form prior to accounting for the analog anti-aliasing filter). The images had 0.5 mm × 0.5 mm in-plane resolution with a matrix size of 384 × 396. To generate "clean" reference images, each multichannel dataset was whitened,[17] then processed with a parameter-free multichannel image denoising method based on the use of wavelet thresholding with Stein's Unbiased Risk Estimator,[30] and was then transformed back to k-space in the original "unwhitened" channel domain. This denoising approach, while imperfect (denoising processes are never perfect and the "clean" images are likely to have lost

some real image features), will at least allow us to mimic having access to thousands of "clean" reference images with matched "noisy" data in a controlled scenario, as needed to test the effects of hidden noise on machine learning methods. To simulate "noisy" data, we then added complex Gaussian noise to the "clean" k-space data following the original interchannel covariance matrix.[17] The noisy k-space data was retrospectively undersampled using a uniform one-dimensional undersampling pattern that kept every fourth line of k-space while also fully sampling the central 24 k-space lines (total acceleration factor = 3.4×). The 3400 slices were partitioned into groups of 3000:200:200 for training, validation, and testing, respectively.

To emulate typical methods comparisons and parameter tuning applications in machine learning, we used the 3000 pairs of fully sampled and undersampled noisy data to train three different machine learning reconstruction approaches: the version of U-Net[31] reconstruction provided alongside the fastMRI dataset,[7] MoDL,[32] and E2E-VarNet.[33] Each reconstruction approach was trained using a range of different parameter variations. The U-Net was fixed to be a nine-layer network with 32 channels in each layer after the first, and was trained with six different numbers of training epochs (100, 110, 120, 130, 140, and 150) and two different loss functions (RMSE and MAE) for a total of 12 different trained U-Net variations. Following the original reference,[32] MoDL was implemented using 10 outer iterations, where each outer iteration comprises a denoising layer (implemented using a five-layer CNN) and a data consistency layer (implemented using a fixed number of 10 conjugate gradient iterations), using the same parameter variations that were used for the U-Net (six different numbers of training epochs and two different loss functions), for a total of 12 different trained MoDL variations. E2EVarNet was implemented with 3 different network variations (a cascade of six U-Nets, where each U-Net uses 20 channels in each layer after the first; a cascade of 7 U-Nets, where each U-Net uses 18 channels in each layer after the first; and a cascade of eight U-Nets, where each U-Net uses 16 channels in each layer after the first), and was trained with two different loss functions (RMSE and MAE) and four different numbers of training epochs (100, 110, 120, and 130) for a total of 24 different trained E2E-VarNet variations. As before, reconstructions obtained with each method were compared against the "clean" and "noisy" rSoS reference images using conventional error metrics (RMSE, MAE, SSIM) and our new NCE metric. Training was performed using the Adam optimizer in all cases, with a learning rate of 0.0003 for U-Net and E2E-VarNet and 0.001 for MoDL.

4 | RESULTS

4.1 | Regularized reconstruction using averaged data

Correlations between reconstruction error metrics obtained with "clean" and "noisy" reference MPRAGE images are shown in Figures 3 and S1 for P-LORAKS (showing behavior as a function of different rank choices and different regularization parameter choices, respectively) and in Figure 4 for SENSE-TV. The trends observed with AC-LORAKS were similar to those observed with P-LORAKS, and are not shown. In most cases (except for SSIM with SENSE-TV), there is relatively poor correlation (correlation coefficient $\rho < 0.55$) between the RMSE, MAE, and SSIM values obtained with the "clean" reference and the "noisy" reference. Importantly, in many of the P-LORAKS cases, there is substantial negative correlation between the error metrics obtained with the "clean" reference and the "noisy" reference, suggesting that if the error metrics from the "noisy" reference are used to tune reconstruction parameters, then this tuning process will be misled by the hidden noise that is present in the "noisy" reference data, which will ultimately lead to an unnecessary degradation in image quality. In contrast, the NCE metric used with the "noisy" reference is much more strongly correlated (correlation coefficient $\rho > 0.9$ in all cases) with the RMSE, MAE, and SSIM values obtained with the "clean" reference.

More insight into this phenomenon can be gained from plotting the same data in a different way. Figure 5 shows the same data from Figure 3 in a different way, with error metrics plotted as a function of the P-LORAKS rank parameter. As can be seen, the optimal rank parameter that would be selected using RMSE, MAE, or SSIM calculated with the "clean" reference (83 or 84 depending on the metric) is often smaller than the optimal rank parameter that would be selected using the "noisy" reference (85, 90, or 91 depending on the metric). The optimal rank that would be selected under the NCE criteria is 84, similar to the choice that would be made using the "clean" reference. Similarly, Figure 6 shows the same data from Figure 4 in a different way, with error metrics plotted as a function of the P-LORAKS regularization parameter. As can be seen, the optimal regularization parameter that would be selected using the "clean" reference (ranging from 0.013 to 0.020 depending on the metric) is much larger than the optimal regularization parameter that would be selected using the "noisy" reference (which is always less than 0.006 for all metrics), while the optimal rank parameters that would be selected based on the NCE criteria (0.013) is more consistent with the values chosen using the "clean" reference.

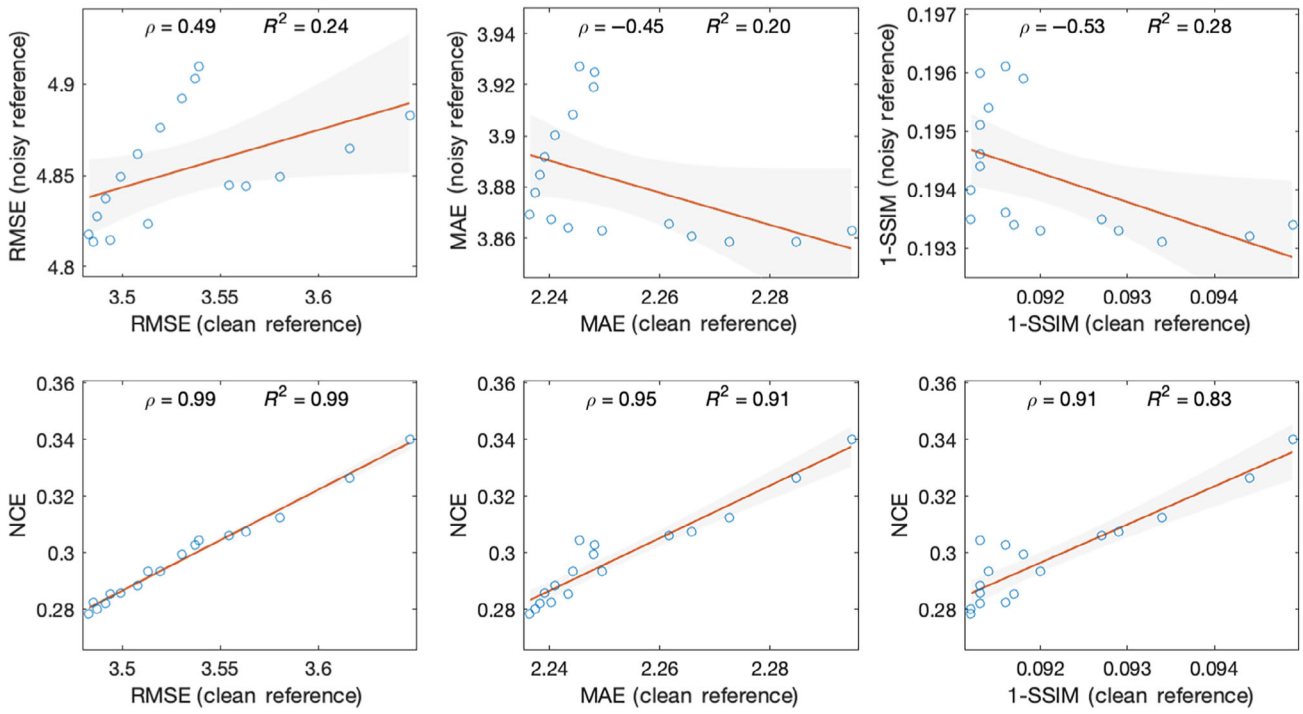

FIGURE 3 Evaluation of P-LORAKS reconstructions (with different rank parameters, for a fixed value of the regularization parameter) using different error metrics. The top row shows correlations between the root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM) values obtained using a high-quality (5×-averaged) "clean" reference and the same metrics obtained using a single-average "noisy" reference, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed noncentral chi error metric.

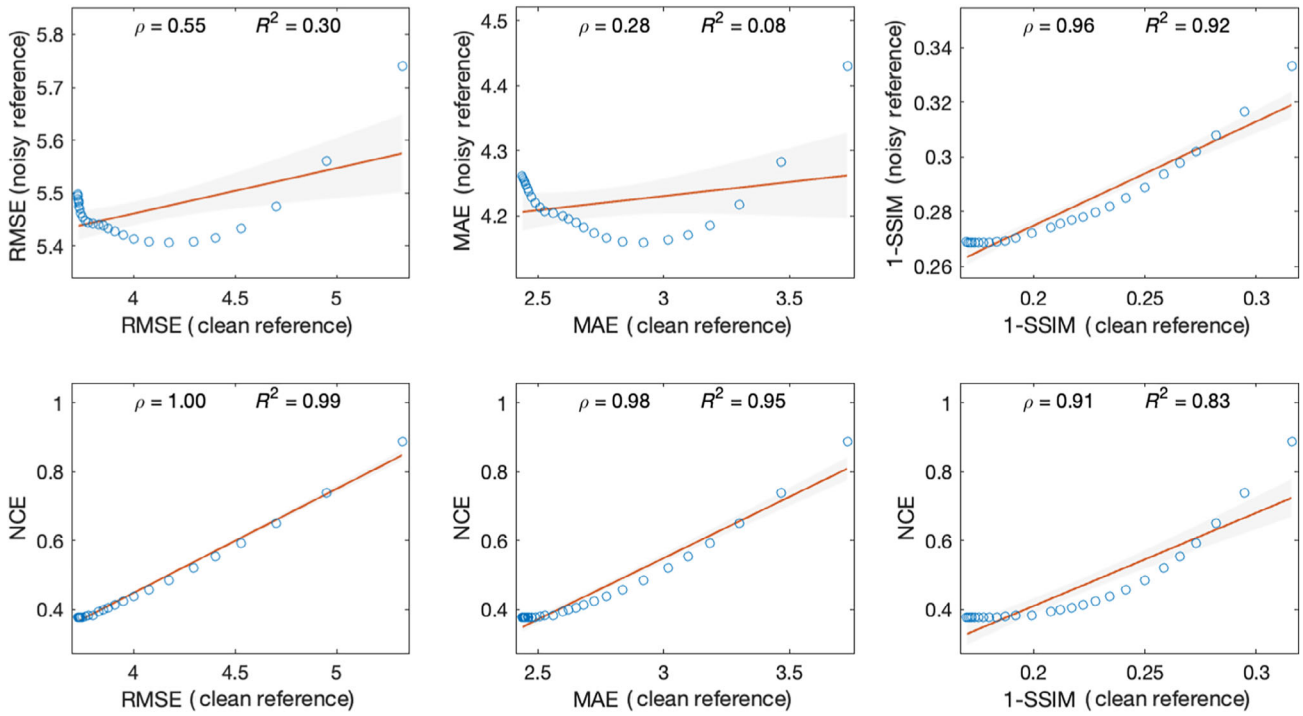

FIGURE 4 Evaluation of SENSE-TV reconstructions (with different regularization parameters) using different error metrics. The top row shows correlations between the root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM) values obtained using a high-quality (5×-averaged) "clean" reference and the same metrics obtained using a single-average "noisy" reference, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed noncentral chi error metric.

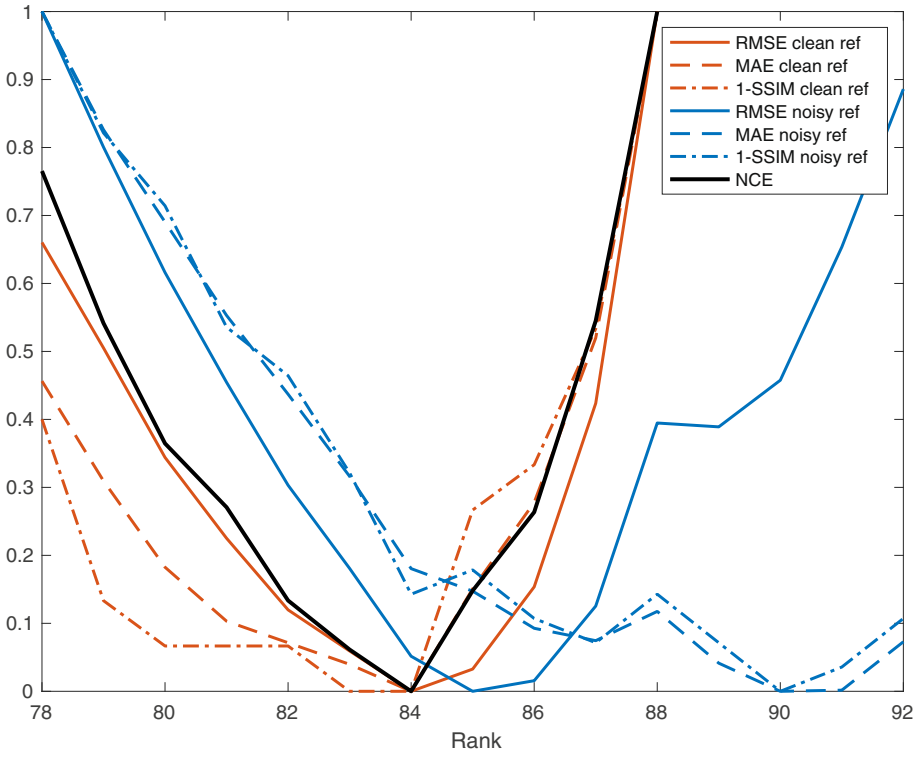

FIGURE 5 The same P-LORAKS results (with different rank parameters) from Figure 3 but now plotted as a function of the rank parameter. Since each metric has a different dynamic range, we have normalized each parameter (so that it ranges from 0 to 1, with smaller values indicating better performance) for easier visualization. Metrics obtained using the "clean" reference are displayed in red, while metrics obtained using the "noisy" reference are displayed in blue, and the noncentral chi error metric is displayed in black.

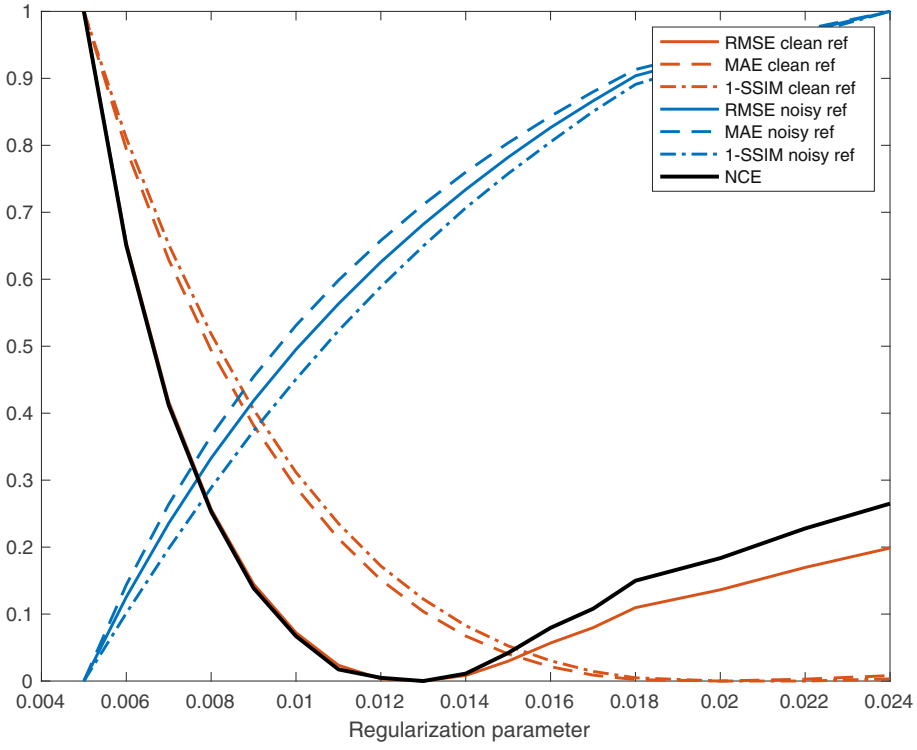

FIGURE 6 The same P-LORAKS results (with different regularization parameters) from Figure 4 but now plotted as a function of the rank parameter. Since each metric has a different dynamic range, we have normalized each parameter (so that it ranges from 0 to 1, with smaller values indicating better performance) for easier visualization. Metrics obtained using the "clean" reference are displayed in red, while metrics obtained using the "noisy" reference are displayed in blue, and the noncentral chi error metric is displayed in black.

Similar results were obtained with SENSE-TV, and are not shown. Collectively, these results further confirm that if a noisy reference image is used for parameter tuning, it is possible for the hidden noise in the reference data to mislead the tuning process when RMSE, MAE, or SSIM are used to judge reconstruction performance, although the same problems are not as severe if the NCE metric is used instead.

To illustrate the potential practical consequences of these hidden noise effects on parameter tuning, Figure 7 shows AC-LORAKS and P-LORAKS images where the parameters in each case were chosen to minimize the

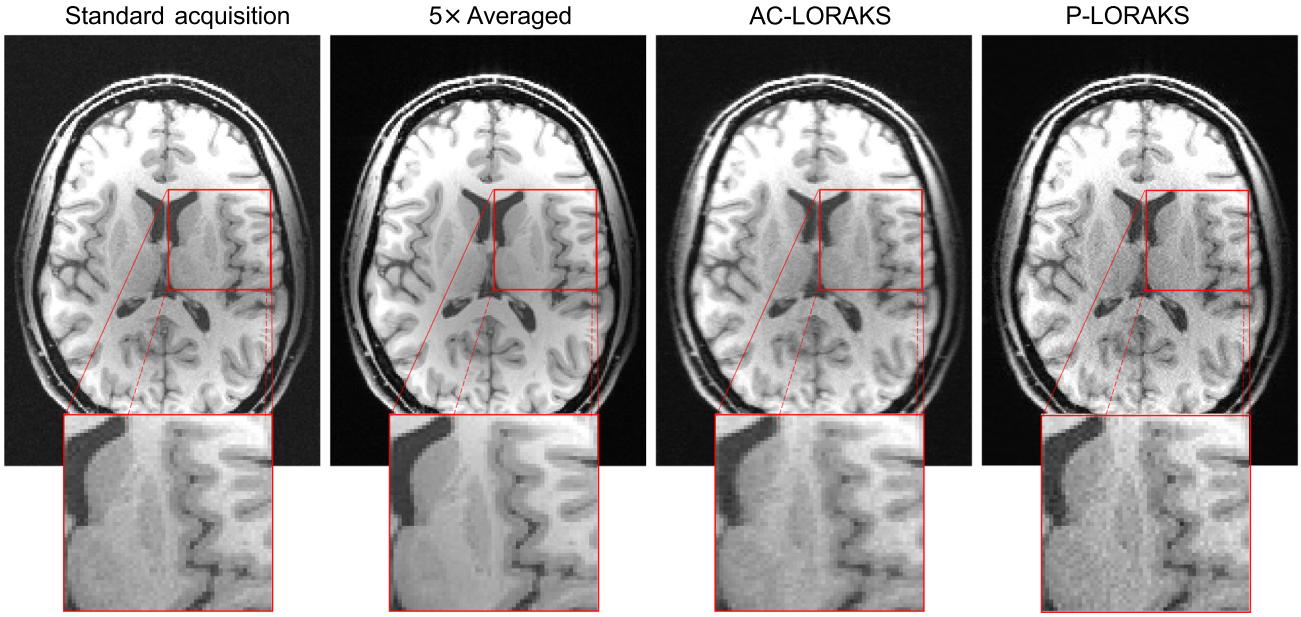

FIGURE 7 Images from the first experiment. The first two images are the noisy and 5×-averaged references. AC-LORAKS achieved better conventional metrics, while P-LORAKS achieved a better non-central chi metric when evaluated with the noisy reference. When compared to the 5×-averaged reference, the P-LORAKS reconstruction outperforms the AC-LORAKS reconstruction with respect to all of the mean-squared error (MSE), mean absolute error (MAE), and the structural similarity index (SSIM) metrics, and is also visually sharper.

RMSE metric calculated using the "noisy" reference image. In this case, the AC-LORAKS reconstruction had better metrics (which were RMSE: 4.07; MAE: 3.00; SSIM: 0.89) than P-LORAKS (which were RMSE: 5.03; MAE: 4.05; SSIM: 0.79) when evaluated using the "noisy" reference image, and would be the method of choice under the conventional parameter-tuning paradigm. However, the AC-LORAKS metrics (which were RMSE: 3.90; MAE: 2.81; SSIM: 0.86) are worse than the P-LORAKS metrics (which were RMSE: 3.56; MAE: 2.24; SSIM: 0.91) with respect to the "clean" reference image, so choosing AC-LORAKS over P-LORAKS in this case would be the wrong decision under these metrics. Qualitatively, while the choice between AC-LORAKS and P-LORAKS is necessarily subjective, we also personally prefer the P-LORAKS result, which appears visually sharper than the AC-LORAKS result. Notably, the NCE metric also prefers the P-LORAKS result (NCE: 0.33) over the AC-LORAKS result (NCE: 2.79), consistent with the metrics obtained with the "clean" reference and also matching our personal preferences.

4.2 | Machine learning reconstruction with denoised data

Figures 8–10 show correlations between reconstruction error metrics obtained with "clean" and "noisy" reference images using the fastMRI data. For the U-Net (Figure 8), we observe that there is modestly good correlation between the RMSE and MAE values corresponding to the "clean" and "noisy" references, although the SSIM values were poorly correlated. In this case, the NCE metric is slightly more correlated with the "clean" RMSE than the "noisy" RMSE was, and the NCE metric is much more correlated with the "clean" SSIM than the "noisy" SSIM was, although is a little worse at correlating with the "clean" MAE than the "noisy" MAE (although these correlations are not too different in this case). For MoDL (Figure 9), we observe that there was good correlation between the RMSE values corresponding to the "clean" and "noisy" references, although the MAE and SSIM values were poorly correlated (with negative correlation for SSIM). In this case, the NCE metric was better correlated with "clean" references than any of the metrics obtained with "noisy" references, with substantial improvements for both MAE and SSIM (although the SSIM correlation is still not particularly good). For E2E-VarNet (Figure 10) there was modest correlation (correlation coefficients between 0.44 and 0.71) between the RMSE, MAE, and SSIM values corresponding to the "clean" and "noisy" references, although the correlation values between the NCE metric and the "clean" reference metrics were nearly perfect (correlation coefficients greater than 0.99 in all cases).

Overall, while there was more variation in this machine learning scenario than there was in the previous regularization scenario, we still observe that there are important cases where the hidden noise present in

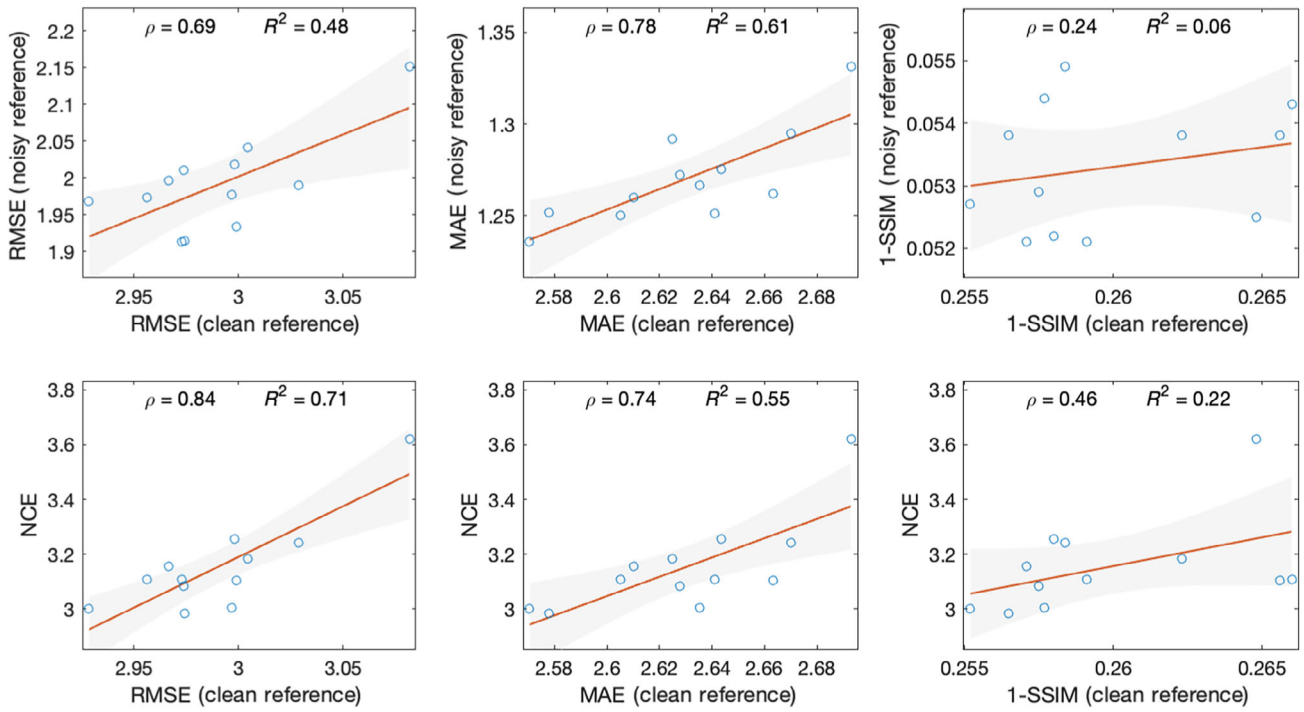

FIGURE 8 Evaluation of U-Net reconstructions using different error metrics. The top row shows correlations between the root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM) values obtained using high-quality (denoised) "clean" references and the same metrics obtained using "noisy" references, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed noncentral chi error metric.

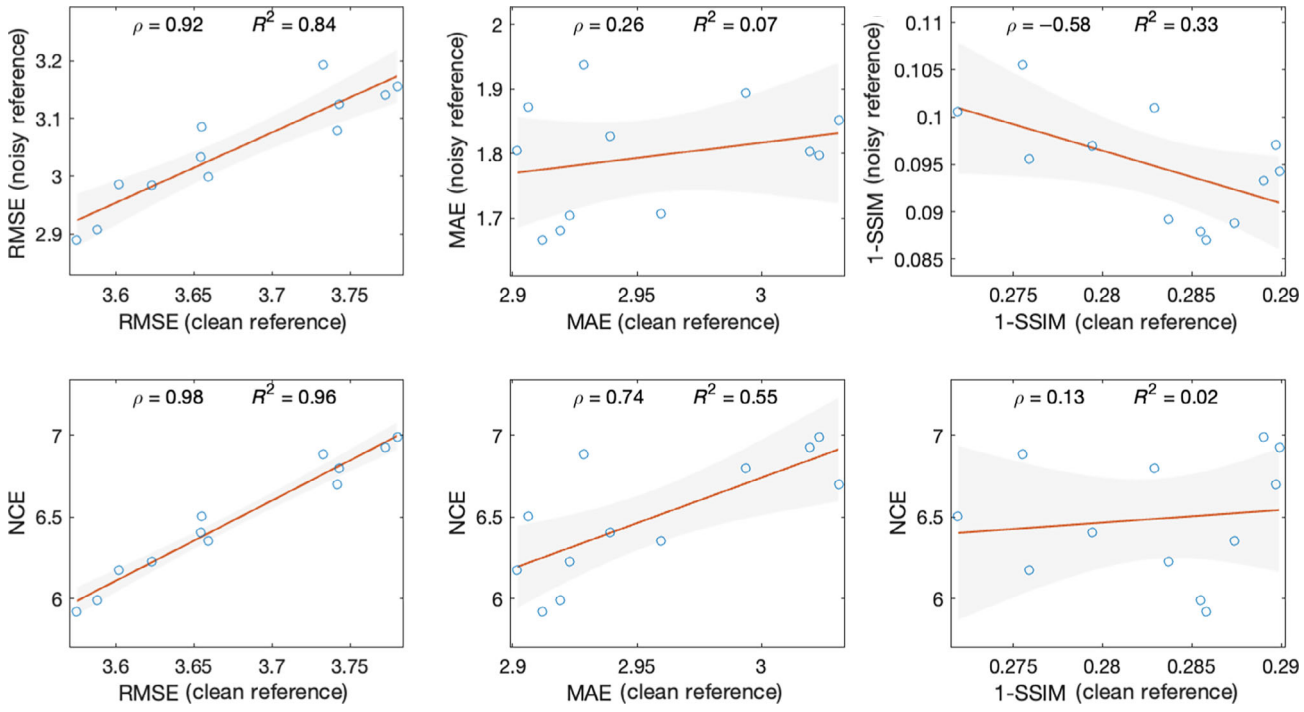

FIGURE 9 Evaluation of MoDL reconstructions using different error metrics. The top row shows correlations between the root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM) values obtained using high-quality (denoised) "clean" references and the same metrics obtained using "noisy" references, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed noncentral chi error metric.

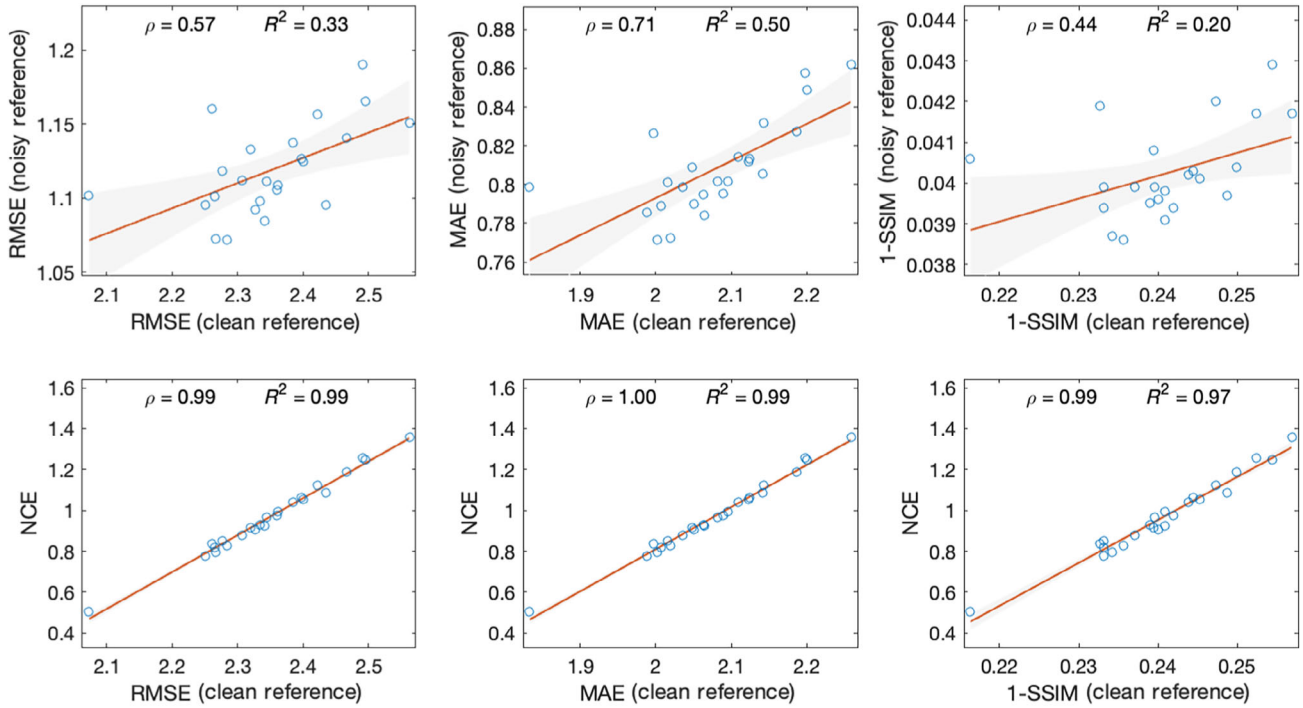

FIGURE 10 Evaluation of E2E-VarNet reconstructions using different error metrics. The top row shows correlations between the root mean-squared error (RMSE), mean absolute error (MAE), and the structural similarity index (SSIM) values obtained using high-quality (denoised) "clean" references and the same metrics obtained using "noisy" references, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed noncentral chi error metric.

noisy reference images is likely to confound the optimal training of machine learning reconstruction methods. In addition, our proposed NCE metric often alleviated this issue.

5 | DISCUSSION AND CONCLUSIONS

Our results demonstrate that, although the effects of hidden noise are usually neglected in the modern image reconstruction literature, this noise can substantially confound the assessment of image reconstruction performance. Under the conventional paradigm, this will lead to suboptimal ranking of different reconstructions and unnecessary degradations in image quality. We believe that awareness of this issue should have major consequences for the development and assessment of image reconstruction methods moving forward, since methods that are reported to have the best RMSE, MAE, and SSIM values may not actually have the best performance once the effects of hidden noise are properly accounted for.

The issues of hidden noise can be somewhat mitigated by using error metrics like NCE that are cognizant of noise in the reference data. In our empirical tests, the correspondence between the NCE metric and "clean" error metrics was comparable or superior (frequently substantially superior) than that of conventional "noisy" RMSE, MAE, or SSIM metrics in every case that we tried, and the NCE was especially strongly correlated with the "clean" RMSE. However, despite generally representing an improvement over "noisy" RMSE, MAE, or SSIM metrics, there were also some reconstructions for which NCE still only had moderate correlation with certain "clean" error metrics. This was particularly true for the SSIM metric with U-Net and MoDL reconstruction. This phenomenon is interesting, and represents room for potential improvements over NCE. Although we can only conjecture at this stage, we speculate that this behavior may have something to do with the fact that the U-NET and MoDL reconstructions were less faithful to the "clean" reference images than the images produced by other methods we tried like E2E-VarNet (for which the NCE was strongly correlated with the "clean" SSIM), causing a lack of consistency between the way that reconstructions were ranked by measures like RMSE and MAE versus measures like SSIM. (Note that the performance of U-Net, MoDL, and E2E-VarNet will be dependent on both hyperparameter choices and the specific datasets that are used for training. We generally followed the same parameter choices that were used in the original

publications, although the performance of all three approaches can likely be improved using different choices, and the fact that E2E-VarNet outperformed the other architectures in our limited evaluation should not be construed as an indication that E2E-VarNet will always outperform U-Net and MoDL).

This study focused on scenarios where reference images are obtained by rSoS coil combination of multichannel images with spatially invariant Gaussian noise, which we believe is the most widely used approach in the modern literature. However, other approaches of forming reference images are also sometimes used that do not induce the NCC distribution. For example, it is possible to use complex-valued sensitivity maps in the coil-combination process, which will produce complex-valued images with zero-mean Gaussian noise and potentially higher SNR than would be achieved from rSoS coil combination.[18] (Note in this case that the use of sensitivity maps will generally result in spatially varying Gaussian noise characteristics). Alternatively, it is also possible to avoid coil-combination entirely, instead measuring error on the original Gaussian-distributed multichannel images with spatially invariant noise. (Note in this case that the individual multichannel images are generally much noisier than coil-combined images [see, e.g., Figure 2], which can present its own confounds). With either of these strategies, measuring errors with respect to complex-valued Gaussian-distributed images may be potentially attractive, since the Gaussian noise will have zero mean with less bias than Rician/NCC magnitude images. For example, Figure S2 shows an evaluation of E2E-VarNet reconstructions where coil-sensitivity maps were used instead of rSoS in the coil-combination procedure to form complex-valued reference images. Comparing against the results obtained with rSoS (cf. Figure 10), the correlation between the "noisy" and "clean" error metrics are much better when using this form of coil combination, and the correlations are even slightly better than those obtained for rSoS reference images with the NCE metric. On the other hand, it is important to keep in mind that while these complex-valued reference images may be less susceptible to hidden noise than rSoS images, they still contain hidden noise and can still be confounded by it. For example, Figures S3 and S4 show results where we are reconstructing the previously described brain MPRAGE dataset, but are now reconstructing Nyquist-sampled data (i.e., these are denoising scenarios, although similar behavior is observed with undersampled data with low acceleration factors). In these cases, the fact that hidden noise is still present in the complex-valued reference images still causes the conventional "noisy" error metrics to be grossly misled, strongly preferring the original noisy dataset over the denoised images that are preferred by the "clean" error metrics. In contrast, the NCE metric with rSoS coil combination is much more strongly correlated with the "clean" error metrics because it properly accounts for the presence of noise. Thus, while hidden noise is of special concern when using rSoS reference images, it can still be an important consideration (i.e., it can present potential confounds for the ranking of reconstruction methods and the interpretation of results) when using other types of reference images.

The observation that NCE frequently has certain advantages over RMSE, MAE, and SSIM suggests that it may be useful when choosing reconstruction parameters or optimizing machine learning reconstruction models. This was already demonstrated for P-LORAKS reconstruction (cf. Figures 5 and 6), where NCE enabled improved selections of regularization and rank parameters. Our preliminary experience using NCE as a loss function suggests that similar improvements could be achieved when training large machine learning models (not shown due to space constraints). Although a thorough exploration is beyond the scope of the present article, we expect that this will be a fruitful direction for further research.

The appropriateness of the NCC model in Equation (4) and the corresponding NCE model (with "effective" $\tilde{\sigma}$ and $\tilde{L}$) in Equation (5) is dependent on forming an rSoS image from multichannel images that each contain spatially-uncorrelated Gaussian noise with spatially invariant characteristics. In this work, we satisfied these assumptions by starting with Nyquist-sampled Cartesian k-space data and forming images through the use of simple linear Fourier inversion, without using other processing steps that might influence the noise distribution. However, it is important to note that there are potential scenarios where these assumptions may be violated, and it is important to approach noise modeling cautiously. For example, our assumption that noise is spatially invariant will likely be violated if images are obtained by applying parallel imaging reconstruction methods to undersampled k-space data, or if sensitivity maps are used to combine multichannel images, or if some form of bias-field/intensity nonuniformity correction is applied, or if some form of interpolation, translation, or other spatial transformation is applied to the data, or if some form of advanced image reconstruction or denoising method is applied. In such cases, obtaining the best possible performance would would likely require modeling the effects of these operations on the noise distribution.[11]

It should also be noted that, while we focused on scalar image quality metrics in this work, image quality can be quite nuanced and multidimensional. Important information that is useful for deciding between two different

reconstructed images can be lost when the quality of an image is summarized by a single number that does not account for different dimensions of image quality such as SNR, contrast, and spatial resolution,[36,37] or different tasks like segmentation or pathology detection that the reconstructed images will be used for downstream.[38] As such, while the scalar NCE metric perhaps has some advantages over other scalar metrics like RMSE, MAE, and SSIM, it still shares the same limitation of all scalar metrics that it cannot possibly provide separate insights into all of the multifold aspects of image quality that may be important in the context of a given MRI application. This means that NCE should still be used with the same amount of caution that should be used with every other scalar performance metric. As a result, we believe that our primary contribution is the observation that hidden noise is an important factor to consider, which has the potential to confound performance evaluation in a wide variety of image reconstruction scenarios.

ENDNOTE

[*] Although the word "noise" is sometimes used in different contexts within MRI (e.g., physiological noise, acoustic noise, etc.), this article always uses "noise" to refer to the random (Johnson–Nyquist) signal fluctuations present in the measured data that result from the thermal agitation of charged particles within the subject and the receiver chain.[6]

[†] It is worth mentioning that the classical Rician noise model for fully sampled single-channel magnitude images[12,13] is a special case of the NCC model, although readers should be cautioned that classical spatially invariant Rician noise modeling is frequently an inadequate representation of the noise characteristics of modern image acquisition and reconstruction methods.[9–11]

[‡] Although tangential to the scope of this article, we should also mention for the sake of completeness that there exist approaches that can also be used to better account for the presence of noise in Gaussian-distributed complex-valued reference images. For example, classical statistical metrics like Stein's unbiased risk estimator[34] could potentially be used to estimate the MSE, although for this to be accurate in the case where coil-combined reference images are obtained using sensitivity maps, it would be important for the Stein's unbiased risk estimator calculations to incorporate information about the spatial variation of the Gaussian noise distribution. Alternatively, the Noise2Noise procedure can also be used in scenarios where multiple noisy observations of the same image are available.[35]

ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01-MH116173, R01-NS074980, and R56-EB034349, the Ming Hsieh Institute for Research on Engineering-Medicine for Cancer, and a USC Annenberg Graduate Fellowship.

ORCID

Jiayang Wang https://orcid.org/0009-0005-4703-5213
Justin P. Haldar https://orcid.org/0000-0002-1838-0211

REFERENCES

1. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600-612.
2. Knoll F, Hammernik K, Zhang C, et al. Deep-learning methods for parallel magnetic resonance imaging reconstruction: a survey of the current approaches, trends, and issues. *IEEE Signal Process Mag*. 2020;37:128-140.
3. Liang D, Cheng J, Ke Z, Ying L. Deep magnetic resonance imaging reconstruction: inverse problems meet neural networks. *IEEE Signal Process Mag*. 2020;37:141-151.
4. Sandino CM, Cheng JY, Chen F, Mardani M, Pauly JM, Vasanawala SS. Compressed sensing: from research to clinical practice with deep neural networks. *IEEE Signal Process Mag*. 2020;37:117-127.
5. Haldar JP, Liang ZP. "Early" constrained reconstruction methods. In: Doneva M, Akcakaya M, Prieto C, eds. *Magnetic Resonance Image Reconstruction: Theory, Methods, and Applications*. Vol 5. Academic Press; 2022:105-125.
6. Pruessmann KP. Sources of noise and limits of SNR. Paper presented at: Proceedings of ISMRM Weekend Educational Courses. 2009.
7. Zbontar J, Knoll F, Sriram A, et al. fastMRI: an open dataset and benchmarks for accelerated MRI. *arXiv:1811.08839*. 2019.
8. Constantinides CD, Atalar E, McVeigh ER. Signal-to-noise measurements in magnitude images from NMR phased arrays. *Magn Reson Med*. 1997;38:852-857.
9. Dietrich O, Raya JG, Reeder SB, Ingrisch M, Reiser MF, Schoenberg SO. Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magn Reson Imaging*. 2008;26:754-762.
10. Aja-Fernandez S, Vegas-Sanchez-Ferrero G. *Statistical Analysis of Noise in MRI: Modeling, Filtering and Estimation*. Springer International Publishing; 2016.
11. Varadarajan D, Haldar JP. A majorize-minimize framework for Rician and non-central chi MR images. *IEEE Trans Med Imaging*. 2015;34:2191-2202.
12. Henkelman RM. Measurement of signal intensities in the presence of noise in MR images. *Med Phys*. 1985;12:232-233.
13. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. *Magn Reson Med*. 1995;34:910-914.
14. Wang J, An D, Haldar JP. The problem of hidden noise in MR image reconstruction. Paper presented at: Proceedings of ISMRM. 2023; Toronto, Canada:4629.
15. Kay SM. *Fundamentals of Statistical Signal Processing*. Estimation Theory. Vol I. Prentice Hall; 1993.

16. Liang ZP, Lauterbur PC. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*. IEEE Press; 2000.
17. Pruessmann KP, Weiger M, Börnert P, Boesiger P. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn Reson Med*. 2001;46:638-651.
18. Roemer PB, Edelstein WA, Hayes CE, Souza SP, Mueller OM. The NMR phased array. *Magn Reson Med*. 1990;16: 192-225.
19. Huang F, Vijayakumar S, Li Y, Hertel S, Duensing GR. A software channel compression technique for faster reconstruction with many channels. *Magn Reson Imaging*. 2008;26: 133-141.
20. Block KT, Uecker M, Frahm J. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. *Magn Reson Med*. 2007;57: 1086-1098.
21. Lustig M, Donoho D, Pauly JM. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn Reson Med*. 2007;58:1182-1195.
22. Haldar JP. Autocalibrated LORAKS for fast constrained MRI reconstruction. Paper presented at: Proceedings of IEEE ISBI, New York, USA. 2015:910-913.
23. Haldar JP. Low-rank modeling of local *k*-space neighborhoods (LORAKS) for constrained MRI. *IEEE Trans Med Imaging*. 2014;33:668-681.
24. Haldar JP, Zhuo J. P-LORAKS: low-rank modeling of local k-space neighborhoods with parallel imaging data. *Magn Reson Med*. 2016;75:1499-1514.
25. Lobos RA, Chan CC, Haldar JP. New theory and faster computations for subspace-based sensitivity map estimation in multichannel MRI. *IEEE Trans Med Imaging*. 2024;43: 286-296.
26. Haldar JP, Hernando D, Liang ZP. Compressed-sensing MRI with random encoding. *IEEE Trans Med Imaging*. 2011;30:893-903.
27. Haldar JP. *Constrained Imaging: Denoising and Sparse Sampling*. Ph.D. Thesis. University of Illinois at Urbana-Champaign. 2011.
28. Kim TH, Haldar JP. *LORAKS Software Version 2.0: Faster Implementation and Enhanced Capabilities*. Technical Report No. USC-SIPI-443. University of Southern California. 2018.
29. Oppenheim AV, Schafer RW. *Discrete-Time Signal Processing*. Prentice Hall; 1999.
30. Luisier F, Blu T. SURE-LET multichannel image denoising: interscale orthonormal wavelet thresholding. *IEEE Trans Image Process*. 2008;17:482-492.
31. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Paper presented at: Proceedings of Medical Image Computing and Computer-Assisted Intervention, Munich, Germany. 2015:234-241.
32. Aggarwal HK, Mani MP, Jacob M. MoDL: model-based deep learning architecture for inverse problems. *IEEE Trans Med Imaging*. 2019;38:394-405.
33. Sriram A, Zbontar J, Murrell T, et al. End-to-end variational networks for accelerated MRI reconstruction. Paper presented at: Proceedings of Medical Image Computing and Computer-Assisted Intervention, Lima, Peru. 2020: 64-73.
34. Stein CM. Estimation of the mean of a multivariate normal distribution. *Ann Stat*. 1981;9:1135-1151.
35. Lehtinen J, Munkberg J, Hasselgren J, et al. Noise2Noise: learning image restoration without clean data. Paper presented at: Proceedings of International Conference on Machine Learning. 2018; Stockholm, Sweden:80.
36. Kim TH, Haldar JP. The Fourier radial error spectrum plot: a more nuanced quantitative evaluation of image reconstruction quality. Paper presented at: Proceedings of IEEE International Symposium on Biomedical Imaging, Washington, D.C, USA. 2018:61-64.
37. Chan CC, Haldar JP. Local perturbation responses and checkerboard tests: characterization tools for nonlinear MRI methods. *Magn Reson Med*. 2021;86:1873-1887.
38. Desai AD, Schmidt AM, Rubin EB, et al. SKM-TEA: a dataset for accelerated MRI reconstruction with dense image labels for quantitative clinical evaluation. Paper presented at: Proceedings of ISMRM. 2022; London, UK:48.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Figure S1. Evaluation of P-LORAKS reconstructions (with different regularization parameters, for a fixed value of the rank parameter) using different error metrics. The top row shows correlations between the RMSE, MAE, and SSIM values obtained using a high-quality (5×-averaged) "clean" reference and the same metrics obtained using a single-average "noisy" reference, while the bottom row demonstrates correlations between the high-quality reference metrics and our proposed NCE metric.

Figure S2. Evaluation of E2E-VarNet reconstructions when complex-valued reference images (coil-combined using sensitivity maps) are used instead of rSoS reference images. The plots show correlations between the RMSE, MAE, and SSIM values obtained using high-quality (denoised) "clean" references and the same metrics obtained using "noisy" references.

Figure S3. Evaluation of SENSE-TV reconstructions of Nyquist-sampled k-space data (with different regularization parameters). The top row shows correlations between the RMSE, MAE, and SSIM values obtained using "noisy" and "clean" complex-valued reference images (coil-combined using sensitivity maps), while the bottom row shows correlations between the RMSE, MAE, and SSIM values obtained from a "clean" rSoS reference image versus the NCE metric for to "noisy" rSoS reference image. In both cases, the "clean" references were obtained from high-quality (5×-averaged) data and the "noisy" references were obtained from single-average data.

Figure S4. Evaluation of P-LORAKS reconstructions of Nyquist-sampled k-space data (with different regularization parameters, for a fixed value of the rank parameter).

The top row shows correlations between the RMSE, MAE, and SSIM values obtained using "noisy" and "clean" complex-valued reference images (multichannel images without coil combination, with error metrics averaged across all channels), while the bottom row shows correlations between the RMSE, MAE, and SSIM values obtained from a "clean" rSoS reference image versus the NCE metric for to "noisy" rSoS reference image. In both cases, the "clean" references were obtained from high-quality (5×-averaged) data and the "noisy" references were obtained from single-average data.

How to cite this article: Wang J, An D, Haldar JP. The "hidden noise" problem in MR image reconstruction. *Magn Reson Med*. 2024;92:982-996. doi: 10.1002/mrm.30100