Reliability and Reproducibility of Neuromelanin-Sensitive Imaging of the Substantia Nigra: A Comparison of Three Different Sequences

Marieke van der Pluijm, MS,[1,2*] Clifford Cassidy, PhD,[3] Melissa Zandstra, MS,[1] Elon Wallert, MD,[1] Kora de Bruin, CNMT,[1] Jan Booij, MD, PhD,[1] Lieuwe de Haan, MD, PhD,[2] Guillermo Horga, MD, PhD,[4] and Elsmarieke van de Giessen, MD, PhD[1]

Background: Neuromelanin-sensitive MRI (NM-MRI) of the substantia nigra provides a noninvasive way to acquire an indirect measure of dopamine functioning. Despite the potential of NM-MRI as a candidate biomarker for dopaminergic pathology, studies about its reproducibility are sparse.
Purpose: To assess the test–retest reproducibility of three commonly used NM-MRI sequences and evaluate three analysis methods.
Study Type: Prospective study.
Population: A total of 11 healthy participants age between 20–27 years.
Field Strength/Sequence: 3.0T; NM-MRI gradient recalled echo (GRE) with magnetization transfer (MT) pulse; NM-MRI turbo spin echo (TSE) with MT pulse; NM-MRI TSE without MT pulse.
Assessment: Participants were scanned twice with a 3-week interval. Manual analysis, threshold analysis, and voxelwise analysis were performed for volume and contrast ratio (CR) measurements.
Statistical Tests: Intraclass correlation coefficients (ICCs) were calculated for test–retest and inter- and intrarater variability.
Results: The GRE sequence achieved the highest contrast and lowest variability (4.9–5.7%) and showed substantial to almost perfect test–retest ICC (0.72–0.90) for CR measurements. For volume measurements, the manual analysis showed a higher variability (10.7–17.9%) and scored lower test–retest ICCs (–0.13–0.73) than the other analysis methods. The threshold analysis showed higher test–retest ICC (0.77) than the manual analysis for the volume measurements.
Data Conclusion: NM-MRI is a highly reproducible measure, especially when using the GRE sequence and CR measurements. Volume measurements appear to be more sensitive to inter/intrarater variability and variability in placement and orientation of the NM-MRI slab. The threshold analysis appears to be the best alternative for volume analysis.
Level of Evidence: 2
Technical Efficacy Stage: 1

J. MAGN. RESON. IMAGING 2021;53:712–721.

IN VIVO VISUALIZATION of the dopamine system is of interest due to its role in a variety of psychiatric and neurodegenerative disorders, including Parkinson's disease (PD)[1] and psychosis.[2] The substantia nigra (SN) in the

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27384

Received Jul 1, 2020, Accepted for publication Sep 18, 2020.

*Address reprint requests to: M.P., MSc, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands. E-mail: m.vanderpluijm@amsterdamumc.nl
Contract grant sponsor: Netherlands Organisation for Health Research and Development (ZonMw); Contract grant number: Veni grant 91618075. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

From the [1]Department of Radiology and Nuclear Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [2]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [3]University of Ottawa Institute of Mental Health Research, affiliated with The Royal, Ottawa, Ontario, Canada; and [4]Department of Psychiatry, New York State Psychiatric Institute, Columbia University Medical Center, New York, New York, USA

mesencephalon is the location from where dopaminergic neurons project to the striatum, forming the nigrostriatal pathway.[3] A novel neuromelanin-sensitive magnetic resonance imaging sequence (NM-MRI) provides a noninvasive way to acquire an indirect measure of dopamine functioning in the SN.[4] NM-MRI has been successfully used to examine changes in the SN in PD and schizophrenia[2,5,6] and seems promising as a biomarker in these disorders. Considering the noninvasive nature of NM-MRI, it has the potential to be applied in clinical practice.

Neuromelanin (NM) is synthesized from cytosolic dopamine and dihydroxyphenylalanine derivatives that have not been taken up into synaptic vesicles.[7,8] After iron-dependent oxidation of the cytosolic dopamine, NM-iron complexes are stacked inside autophagic organelles that fuse with lysosomes, and lipid and protein components forming the final autolysosomal NM-containing organelles.[7,9] These NM-containing organelles accumulate over age in the SN,[10] or rather show an inverted U-shaped age effect.[11] The paramagnetic NM-iron complexes lead to T_1 reduction, which contributes to the NM-MRI contrast.[4]

NM-sensitive T_1-weighted turbo spin echo (TSE) is the most frequently used NM-MRI sequence and multiple studies have shown a reliable decrease of NM-MRI signal in the SN of patients with PD.[5,12-27] The TSE sequence can be performed with or without a magnetization transfer (MT) pulse. An off-resonance MT pulse suppresses the contribution of macromolecules to the signal and can thereby increase contrast.[4,24,28,29] A smaller number of patient studies have been performed with an NM-sensitive gradient echo pulse (GRE) sequence with an MT pulse.[6,30]

Despite the promise of NM-MRI as a biomarker, studies on the reproducibility and reliability of the different sequences have been sparse. Reproducibility provides vital information for study designs, since outcome measures with lower reliability have diminished power. One study has shown a lower sensitivity for imaging the SN with a TSE sequence compared to a GRE sequence,[31] but reproducibility was not compared. Reproducibility studies have been performed for the GRE sequence and yielded excellent results for volume measurements[32] and contrast ratio measurements,[6,32,33] while using the TSE sequence, a study has shown moderate reproducibility in the noradrenaline-rich locus coeruleus.[34] A study directly comparing the GRE and TSE NM-MRI sequences in terms of their reproducibility in SN imaging would be useful.

In addition to differences in image acquisition, there are also differences in the analysis of NM-MRI data. Most studies have used an average contrast ratio based on a manual approach in which the SN and a reference region are manually traced on one or more axial slices in each NM-MRI scan. While this has shown differences between groups,[15,19,20,24,35] it does not assess the entire SN. Furthermore, this method is sensitive to inter/intrarater variability resulting from variability in the placement and orientation of the NM-MRI imaging slab within and across studies. An intensity threshold method avoids the inter/intrarater variability and when applied to scans normalized to Montreal Neurological Institute (MNI) space it can give an estimate of the entire SN, independent of slab placement.[32] More recently, Cassidy et al validated a voxelwise approach.[6] This method is semiautomated using an average mask of the normalized dataset instead of a subject-specific mask. This approach captures the voxel anatomical information in the scan and can be implemented for various measurements, including mapping the (sub)regional variation of the SN.

In order to further develop NM-MRI for research and clinical applications, it is important to determine optimal acquisition and analysis methods. Therefore, the aim of this study was to compare the test–retest reproducibility of three NM-MRI sequences in SN imaging; 1) the GRE sequence with MT pulse, 2) the TSE sequence with MT pulse, and 3) the TSE sequence without MT pulse and, also, to assess and compare the reliability of three analysis techniques; i) manual analysis, ii) threshold analysis, and iii) voxelwise analysis.

Materials and Methods

Participants

This study was approved by the local Medical Ethics Committee. All participants gave written informed consent prior to the first scan after the nature of the procedure had been fully explained. Eleven healthy participants (mean [SD] age: 24.82 [2.04], range: 20–27 years, seven male and four female) were included in the study. Prior to inclusion, all participants were screened by means of an interview and excluded if they had a history of neurological or psychiatric disorders, used any medication (with the exception of contraceptives), or had any MRI contraindications.

Image Acquisition

All MR data were acquired at a single center using a 3T Ingenia MRI system (Philips, Best, The Netherlands) with a 32-channel SENSE head coil. All participants participated in two identical NM-MRI scanning sessions, with a 3-week interval (mean [SD] days: 20.9 [1.4], range: 18–24 days).

For slice placement and registration, transversal high-resolution structural T_1-weighted volumetric images, with full head coverage, were acquired (echo time [TE] / repetition time [TR] = 4.1/9.0 msec; 189 slices; field of view [FOV] = 284 × 284 × 170 mm; voxel size: 0.9 × 0.9 × 0.9 mm, flip angle [FA] = 8°). On these, the NM-MRI sections were placed perpendicular to the fourth ventricle floor with coverage from the posterior commissure to halfway through the pons.

The following three NM-MRI scans were acquired; 1) GRE MToff-resonance pulse (GRE MT on) (TE/TR = 3.9/260 msec, FA = 40°, 8 slices, slice thickness = 2.5 mm, in-plane resolution = 0.39 × 0.39 mm^2, FOV = 162 × 199 mm, number of signal averages [NSA] = 2, magnetization transfer frequency offset = 1200 Hz and duration = 15.6 msec, based on[6,33]); 2) TSE with

MT off-resonance pulse (TSE MT on) (TE/TR = 10/641 msec, FA = 90°, 8 slices, slice thickness = 2.5 mm, in-plane resolution = 0.40 × 0.40 mm^2, FOV = 180 × 180 mm, NSA = 2, magnetization transfer frequency offset = 1200 Hz and duration = 15.6 msec, based on[24]); 3) TSE without MT pulse (TSE MT off) (TE/TR = 10/500 msec, FA = 90°, 8 slices, slice thickness = 2.5 mm, in-plane resolution = 0.40 × 0.40 mm^2, FOV = 180 × 180 mm, NSA = 2, based on[20]).

Manual Segmentation

ITK-Snap (v. 3.6.0, www.itksnap.org)[36] was used to manually segment the SN. In addition, the crus cerebri (CC) and red nucleus (RN) were segmented and served as reference areas.[6,20,33] Segmentation was performed by three independent raters (M.Z., K.B., E.W.) and the raters segmented both the test and retest scans twice. The interval between segmentation 1 and segmentation 2 was a minimum of 3 weeks and a maximum of 6 weeks. To ensure raters had the same segmentation approach, a segmentation protocol was used and all attended a training session. No raters had experience with the NM-MRI segmentations prior to this study. For every sequence two contrast ratios (CR), $CR_{SN\text{-}RN}$ and $CR_{SN\text{-}CC}$, were determined. These were calculated as: $CR_{SN\text{-}RN} = (S_{SN} - S_{RN})/S_{RN}$ and $CR_{SN\text{-}CC} = (S_{SN} - S_{CC})/S_{CC}$, where S_{SN}, S_{RN}, and S_{CC} represent the mean signal intensities of the SN, RN, and CC, respectively. For each participant the two slices with the highest voxel intensity were segmented. Segmentation of the CC consisted of six default circles (three on each side of the SN), each with a diameter of 8 mm.

Standardized Preprocessing

For the standardized analyses, we preprocessed the NM-MRI scans using Statistical Parametric Mapping's (SPM 12; Wellcome Trust, London, UK). We first coregistered the NM-MRI retest scans to the NM-MRI test scans and subsequently coregistered both to the T_1-weighted test scans. Tissue segmentation was performed using the T_1-weighted test scan. All scans were normalized into MNI standard brain space using DARTEL routines with a gray and white matter template generated from all T_1-weighted test scans and spatially smoothed with a 1-mm full-width at half-maximum Gaussian kernel. For post-hoc analysis, the preprocessing was performed without spatial smoothing to assess the effect of spatial smoothing. All images were visually inspected following each preprocessing step.

Semiautomated Thresholding Segmentation

For each sequence a large area around the (left and right) SN was manually traced on the standardized average image with ITK-Snap. This was done carefully to avoid contamination from CSF space. This mask was overlaid on the individual NM-MRI scans in MNI space and voxels with signal intensity S_v >3 standard deviations from S_{CC} were considered part of the SN. All high-intensity voxels generated by the thresholding method were visually inspected to ensure no outlying/aberrant voxels were included in the mask.

Voxelwise Analysis

We used FSL (FMRIB Software Library, v. 5.0.10, Oxford University, UK) to create one standardized average for each of the three sequences based on the 22 standardized NM-MRI scans (test and retest). Template population masks for both the SN and CC were created for each sequence by manual tracing with ITK-Snap on the standardized average image (Fig. 1b). For each scan and voxel in the SN mask a CR_v was calculated as $CR_V = (S_v - S_{CC})/S_{CC}$. Voxels with a CR_V smaller than 0 or greater than 3 standard deviations from the mean were excluded.

Statistical Analysis

To assess reliability in test–retest, intrarater, and interrater reliability the intraclass correlation coefficient (ICC) was used.[37] For the manual analysis the ICC estimates and their 95% confident intervals (CIs) were calculated using SPSS 26.0 (IBM, Armonk, NY) and the ICC for the thresholding and voxelwise analyses were calculated using MatLab (MathWorks, R2016a, Natick, MA). The test–retest ICC was based on single measures and consistency two-way mixed-effects and the ICC for the intrarater and interrater agreement was based on single measures and an absolute agreement two-way mixed-effects model. Standard thresholds were used for interpretability of ICC values: "almost perfect" for ICC 0.81–1.00, "substantial" for ICC 0.61–0.80, "moderate" for ICC between 0.41–0.60, "fair" for ICC 0.21–0.40, "slight" 0.00–0.20, and "poor" for ICC <0.00.[37]

Test–retest variability was assessed as a measure of agreement and was calculated using the following equation: $VAR = \frac{|test - retest|}{(test + retest)/2} \cdot 100\%$ and performed for the manual analysis on both contrast ratios ($CR_{SN\text{-}RN}$ and $CR_{SN\text{-}CC}$) and SN volumes for all three sequences and raters. Furthermore, Bland–Altman plots for test and retest were constructed as an additional measure of agreement for the manual analysis.

For the semiautomated thresholding segmentation, additionally the Dice similarity coefficient (DSC) was calculated to determine reproducibility between the mean test and retest volume measurements. The DSC was calculated with MatLab and defined as DSC = (2*volume (Test ∩ Retest)) / (volume(Test) + volume (Retest)), where ∩ represents the intersection operator.

Results

Manual Segmentation

In all subjects the SN was consistently detected as an area of hyperintensity (Fig. 1a). Tables 1 and 2 show the test–retest variability, test–retest reliability, intrarater reliability, and interrater reliability based on the manual segmentation protocol for the CR and volume measurements, respectively. For CR, variability was lowest for the GRE MT on sequence with $CR_{SN\text{-}CC}$ analysis. Test–retest ICC was substantial to almost perfect (0.60–0.86) for all three sequences with $CR_{SN\text{-}CC}$ analysis. Also, intrarater ICC was substantial to almost perfect (0.75–0.97) for all three sequences with $CR_{SN\text{-}CC}$ analysis, whereas the interrater ICC was substantial (0.63–0.81). Since the $CR_{SN\text{-}CC}$ yielded better reproducibility than the $CR_{SN\text{-}RN}$, further analysis focused solely on the $CR_{SN\text{-}CC}$ (CR_{SN}). For volume measurements, the TSE MT on sequence yielded the lowest variability (10.74–11.47%) and highest ICCs with a slight to substantial test–retest ICC (0.11–0.73) and intra- and interrater ICC varying from slight to almost perfect (0.34–0.87). Bland–Altman plots of the CR_{SN}, depicted in Fig. 2, give a graphical representation of the agreement for

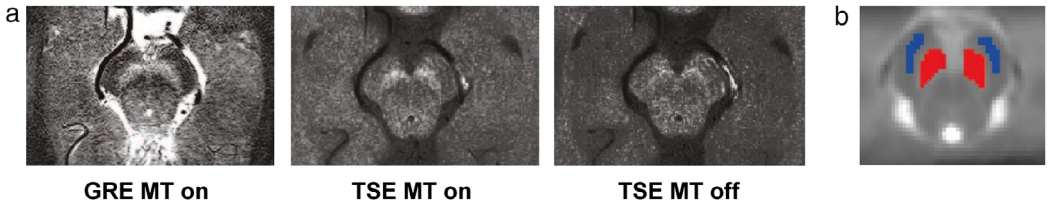

Figure 1: NM-MRI of the substantia nigra. (a) An individual example of the three NM-MRI sequences. (b) Manual segmentation of subsantia nigra (SN) and crus cerebri (CC) mask on a standardized image in MNI space. The mask of the SN is shown in red and the mask of the CC in blue. GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

the CR_{SN} and volume between the test and the retest for all three sequences.

Semiautomated Thresholding Segmentation

The semiautomated thresholding segmented SN volumes were reproducible with a substantial ICC reliability (0.67–0.77, see Table 3). In addition, the semiautomated thresholding segmented SN volume showed significant overlap between the two scans (Table 3), especially for the GRE MT on (0.91). The post-hoc nonsmoothed preprocessed data yielded lower reliability compared to the smoothed preprocessed data (Table 3).

Voxelwise Analysis

The standardized mask did not fit the retest scan of one participant, due to suboptimal imaging slab placement (top part of SN was missing); therefore, this participant was excluded from the automatic analysis.

CR for each voxel in the standardized average was calculated for test and retest (Fig. 3). Two-way mixed, single score ICC and consistency values between test and retest per voxel were calculated, creating a map of ICC values in the SN (Fig. 3). The mean ICC across voxels for each sequence was calculated with the available data in at least 10 participants (Table 3). The results yielded a substantial ICC (0.72) for the GRE sequence, moderate ICC (0.52) for the TSE MT on and fair ICC (0.37) for the TSE MT off. In addition, the ICC for the average CR_{SN} in the whole ICC mask was calculated for each sequence (Table 3). ICC values of the average CR in the SN were almost perfect for the GRE sequence (0.90), substantial for TSE MT on (0.79), and the TSE MT off (0.66). The post-hoc nonsmoothed preprocessed data yielded lower reliability compared to the smoothed preprocessed data (Table 3).

Discussion

This study compared NM-MRI sequences with regard to their reproducibility and evaluated different analysis methods in young healthy controls. Overall, the GRE MT on sequence achieved the best reproducibility and reliability for all analysis methods (manual, threshold, and voxelwise). For CR measurements, the CC was a more stable reference region than the RN and the test–retest, intra- and interrater ICCs ranged from substantial to almost excellent. The SN volume measurements were more variable, with poor to slight test–retest variability and slight to almost perfect inter- and intrarater ICCs for the manual segmention and a substantial ICC for the threshold analysis.

The better performance of the GRE MT on sequence is an important result, as most NM-MRI studies have been performed with TSE sequences.[5,12-27] Our results for the GRE MT on sequence are in line with previous reliability studies. Langley et al found a lower ICC (0.81) for the CR_{SN} for the threshold analysis, even though in our study the two MRI scans were separated by a 3-week interval instead of a single session day.[32] This difference might be due to variation in analysis design. For example, this study used SPM with DARTEL routines for normalization to MNI space, whereas Langley et al used FSL.[32] However, they found a higher ICC (0.94) for the volume measurements using a semiautomated thresholding method, although they yielded a lower DSC (0.80) for the volume measurement. The ICC is more clinically relevant than the DSC, however, since it measures the reproducibility of CR and volume measurements instead of the reproducibility of the volume location and center. The current study replicated the analysis design of Cassidy et al and, indeed, their results (an ICC of 0.95 for the CR_{SN}) are more comparable to ours (0.90), while they had only an hour interval between test and retest acquisitions.[6] The voxelwise analysis also showed similar results, with a mean ICC CR_v of 0.72 compared to a median ICC of 0.64. This result suggests that a longer period between test and retest (3 weeks instead of an hour) does not result in increased variability. A recent study compared different acquisition parameters for GRE MT on sequences and different voxelwise analysis toolboxes[33] with higher ICCs than ours, which is most likely related to further optimized slab placement and coregistration methods. These comparisons emphasize the influence of analysis and acquisition design.

In this study a single measures and absolute agreement two-way mixed-effects model was chosen for the intrarater and interrater reliability, since the purpose was to compare the absolute score from the raters on the same measurement (scan). For the test–retest ICC, single measures and consistency two-way mixed-effects were used. This ICC does not penalize systematic variability across test and retest (eg, if the

TABLE 1. Reproducibility and Reliability Based on the Manual Segmented CR_{SN} of NM-MRI Sequences

	GRE MT on		TSE MT on		TSE MT off	
	$CR_{SN\text{-}RN}$	$CR_{SN\text{-}CC}$	$CR_{SN\text{-}RN}$	$CR_{SN\text{-}CC}$	$CR_{SN\text{-}RN}$	$CR_{SN\text{-}CC}$
Mean CR% (SD)						
Rater 1 (test)	21.59 (2.26)	21.64 (2.13)	21.35 (2.32)	13.51 (2.20)	16.49 (2.96)	9.04 (2.20)
Rater 1 (retest)	22.89 (2.69)	21.42 (2.15)	20.90 (2.57)	12.96 (2.07	16.27 (3.54)	9.14 (2.13)
Rater 2 (test)	20.86 (3.21)	22.48 (2.18)	19.63 (3.45)	13.19 (2.25)	16.29 (3.42)	9.07 (2.21)
Rater 2 (retest)	23.14 (2.80)	21.75 (2.35)	21.13 (2.15)	12.87 (1.76)	16.76 (2.68)	8.98 (2.28)
Rater 3 (test)	21.59 (2.86)	19.72 (1.93)	20.37 (3.05)	12.14 (1.98)	16.54 (2.42)	7.73 (1.72)
Rater 3 (retest)	22.28 (2.81)	20.37 (2.10)	20.63 (1.91)	11.78 (2.01)	16.69 (2.29)	7.94 (1.88)
Raters combined (test)	21.35 (2.74)	21.28 (2.33)	20.45 (2.97)	12.95 (2.16)	16.44 (2.87)	8.61 (2.09)
Raters combined (retest)	22.77 (2.71)	21.18 (2.22)	20.89 (2.17)	12.54 (1.97)	16.58 (2.80)	8.69 (2.11)
Test–retest variability (SD)						
Rater 1	7.47 (5.94)	4.92 (2.58)	7.35 (5.55)	9.92 (7.38)	12.83 (9.09)	8.70 (9.23)
Rater 2	13.51 (9.72)	5.98 (4.25)	11.25 (11.30)	7.41 (5.30)	14.87 (12.38)	15.68 (15.57)
Rater 3	7.86 (6.81)	5.73 (4.23)	8.65 (9.25)	10.29 (9.18)	8.84 (5.42)	16.55 (9.39)
Test–retest ICC (95% CI)						
Rater 1	0.74 (0.28–0.91)	0.84 (0.51–0.95)	0.70 (0.21–0.91)	0.76 (0.32–0.93)	0.73 (0.27–0.92)	0.86 (0.55–0.96)
Rater 2	0.58 (0.05–0.87)	0.79 (0.40–0.94)	0.61 (0.05–0.88)	0.85 (0.54–0.96)	0.54 (–0.05–0.85)	0.60 (0.04–0.88)
Rater 3	0.69 (0.19–0.91)	0.79 (0.39–0.94)	0.55 (–0.04–0.84)	0.67 (0.15–0.90)	0.75 (0.320–0.93)	0.66 (0.13–0.89)
Rater ICC (95% CI)						
Intrarater ICC (R1)	0.94 (0.86–0.98)	0.92 (0.80–0.97)	0.80 (0.58–0.91)	0.97 (0.92–0.99)	0.86 (0.70–0.94)	0.85 (0.68–0.94)
Intrarater ICC (R2)	0.85 (0.68–0.94)	0.88 (0.56–0.96)	0.75 (0.46–0.89)	0.75 (0.49–0.89)	0.81 (0.58–0.92)	0.87 (0.66–0.95)
Intrarater ICC (R3)	0.93 (0.83–0.97)	0.93 (0.83–0.97)	0.55 (0.17–0.78)	0.90 (0.79–0.96)	0.86 (0.69–0.94)	0.92 (0.72–0.97)
Interrater ICC	0.75 (0.57–0.88)	0.63 (0.22–0.84)	0.65 (0.44–0.82)	0.81 (0.45–0.93)	0.62 (0.39–0.81)	0.79 (0.44–0.92)

Given are the test–retest variability for each rater and ICC values with 95% confidence interval.
SN: substantia nigra; CR: contrast ratio; CR%: contrast ratio * 100; SD: standard deviation; R1: rater 1; R2: rater 2; R3: rater 3; ICC: intraclass correlation coefficient; 95% CI: 95 percent confidence interval; GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

TABLE 2. Reproducibility and Reliability Based on the Manual Segmented SN Volumes of the NM-MRI Sequences

	GRE MT on	TSE MT on	TSE MT off
Mean volume mm^3 (SD)			
Rater 1 (test)	355.52 (54.53)	337.93 (64.73)	328.15 (69.43)
Rater 1 (retest)	373.91 (56.48)	362.99 (67.62)	306.11 (48.99)
Rater 2 (test)	333.87 (35.48)	343.10 (55.18)	302.84 (42.80)
Rater 2 (retest)	308.35 (36.73)	340.14 (45.73)	304.31 (37.37)
Rater 3 (test)	365.56 (52.83)	386.50 (42.14)	378.63 (36.38)
Rater 3 (retest)	391.11 (40.90)	414.67 (40.96)	380.67 (47.00)
Raters combined (test)	351.65 (48.73)	355.85 (57.49)	336.54 (59.30)
Raters combined (retest)	357.79 (57.03)	372.50 (60.08)	330.36 (56.41)
Test–retest variability (SD)			
Rater 1	14.31% (13.08)	10.74% (10.16)	13.03% (12.76)
Rater 2	13.04% (10.68)	10.99% (7.84)	17.88% (9.82)
Rater 3	14.47% (12.05)	11.47% (9.84)	12.76% (7.83)
Test–retest ICC (95% CI)			
Rater 1	0.13 (–0.48–0.66)	0.73 (0.26–0.92)	0.43 (–0.19–0.81)
Rater 2	0.05 (–0.54–0.61)	0.57 (–0.11–0.86)	–0.13 (–0.66–0.48)
Rater 3	–0.10 (–0.64–0.51)	0.11 (–0.50–0.65)	–0.01 (–0.59–0.57)
Rater ICC (95% CI)			
Intrarater ICC (R1)	0.84 (0.67–0.93)	0.87 (0.71–0.94)	0.64 (0.32–0.83)
Intrarater ICC (R2)	0.33 (–0.9–0.65)	0.35 (–0.9–0.67)	0.53 (0.17–0.77)
Intrarater ICC (R3)	0.82 (0.30–0.94)	0.72 (0.45–0.87)	0.58 (0.12–0.81)
Interrater ICC	0.11 (–0.07–0.36)	0.34 (0.07–0.61)	0.21 (–0.01–0.47)

Given are the test–retest variability for each rater and ICC values with 95% confidence interval.
R1: rater 1; R2: rater 2; R3: rater 3; SD: standard deviation; ICC: intraclass correlation coefficient; 95% CI: 95% confidence interval; GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

retest is consistently higher than the test measurements). The consistency ICC has been used in previous reliability studies of the GRE MT on NM-MRI.[6,33] We observed, however, hardly any differences between the absolute agreement and consistency ICC values.

A previous study also compared different NM-MRI sequences and showed a higher CR for the GRE sequence compared to a TSE sequence.[31] Our study replicated this finding and additionally indicated that a higher contrast goes together with better reproducibility. Adding the MT pulse increased the contrast for the TSE sequences and also resulted in an overall better reproducibility. This supports previous work that has shown that contrast in NM-MRI is associated with MT effects (next to T_1 reduction by neuromelanin), which is probably partly related to higher macromolecular content in the adjacent white matter than in the gray matter of the SN.[4,28,38]

It is reassuring to see that different analysis methods achieved substantial to almost perfect reliability for the CR. The method with the highest ICC was the analysis in which all scans were normalized to standard space, before calculating the CR for the whole SN mask. The choice of analysis method may be based on a number of considerations. Manual segmentation might be considered with a small sample size. Since it does not require a normalization step, there is no introduction of normalization errors and the low complexity could increase clinical applicability. In addition, manual segmentation might be more suitable for comparing two

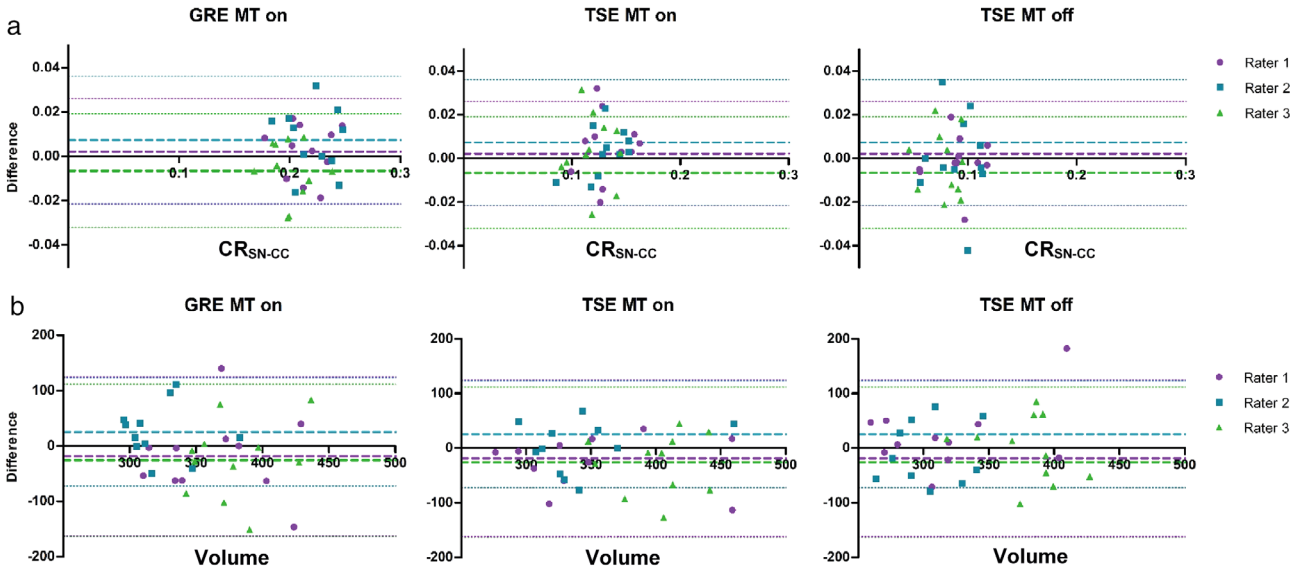

Figure 2: Bland–Altman plots representing the difference between the test and the retest of (a) manual segmented contrast ratios and (b) manual segmented volumes. With the representation of the mean difference (dashed lines) and the limits of agreement (dotted lines), from –1.96 SD to +1.96 SD; in purple Rater 1, in blue Rater 2, in green Rater 3. GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

anatomically different groups (eg, due to atrophy), since this could complicate the registration/normalization process and defining the template mask. However, manual segmentation is labor-intensive and susceptible to rater differences and imaging slab placement, especially for the volume measurements. This is an important finding, since numerous studies

TABLE 3. Mean Test–Retest Reliability of the Semiautomated Analyses With the Contrast Ratios From the Voxelwise Analysis and the Volumes From the Semiautomated Thresholding Segmentation

	GRE MT on	TSE MT on	TSE MT off
Threshold analysis			
ICC Volume (95% CI)	0.77 (0.31–0.94)	0.71 (0.18–0.92)	0.67 (0.12–0.91)
DSC Volume (SD)	0.91 (0.03)	0.71 (0.07)	0.68 (0.13)
Voxelwise analysis			
ICC CR_V (95% CI)	0.72 (0.25–0.92)	0.52 (–0.05–0.84)	0.37 (–0.25–0.78)
ICC CR_{SN} (95% CI)	0.90 (0.66–0.97)	0.79 (0.36–0.94)	0.66 (0.09–0.90)
Analysis without spatial smoothing			
	GRE MT on	**TSE MT on**	**TSE MT off**
Threshold analysis			
ICC Volume (95% CI)	0.78 (0.26–0.94)	0.64 (0.07–0.90)	0.65 (0.08–0.90)
DSC Volume (SD)	0.86 (0.03)	0.72 (0.08)	0.51 (0.18)
Voxelwise analysis			
ICC CR_V (95% CI)	0.63 (–0.00–0.87)	0.38 (–0.21–0.79)	0.26 (–0.34–0.71)
ICC CR_{SN} (95% CI)	0.83 (–0.45–0.95)	0.61 (0.01–0.89)	0.64 (0.06–0.90)

ICC: intraclass correlation coefficient; ICC CR_v: average ICC of voxels in substantia nigra ROI for contrast ratio; ICC CR_{sn}: ICC for average contrast ratio in the substantia nigra ROI; DSC: Dice similarity coefficient; GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

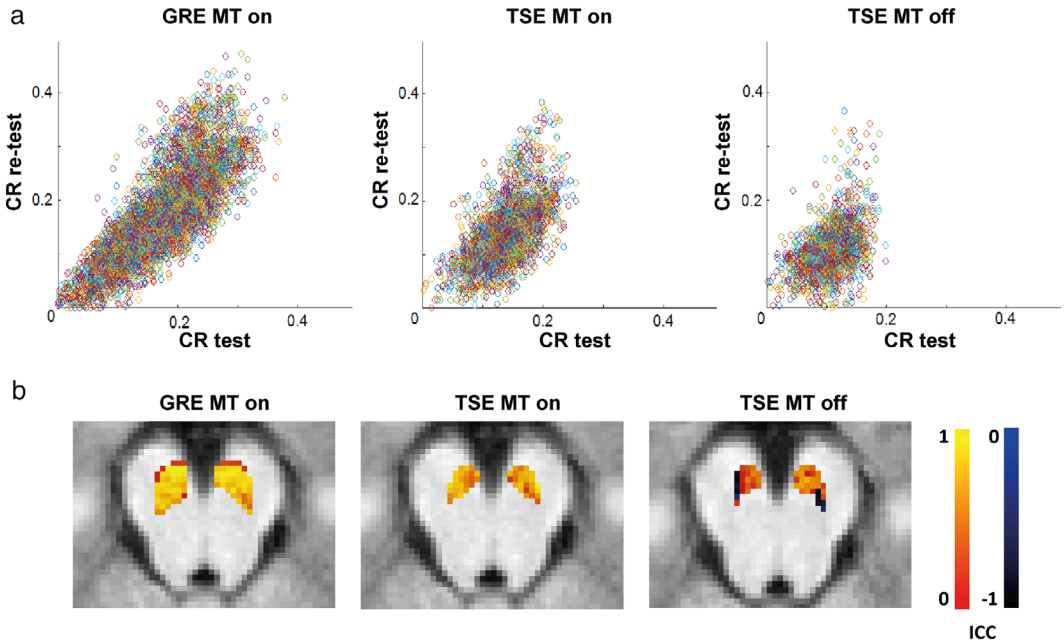

Figure 3: Contrast ratios (a) and ICC values (b) per voxel between test and retest measurement in the substantia nigra for the three different NM-sequences. CR: contrast ratio; GRE: gradient recalled echo; MT: magnetization transfer pulse; TSE: turbo spin echo.

on PD have used the manual approach for volume measurements.[5,14,19,21,22,24] For the manual volume measurements, the test–retest ICCs were mainly poor to slight and considerably lower than the intrarater ICCs, which ranged from slight to excellent. For the CR measurements, however, test–retest ICCs were fair to good, while intrarater ICCs were all excellent. This may indicate that the manual volume measurements were more susceptible to differences in imaging slab placement between test and retest sessions. The thresholding method could be a better alternative for volume measurements and could also be applied without the normalization step. It is less susceptible to rater differences than the manual volume analysis, since is detects the most intense voxels in a large region of interest (ROI) around the SN, but variability from imaging slab placement (in particular, angle) would remain. The normalization step does seem to reduce the test–retest variability, and thus reduce variability from imaging slab placement, with a substantial ICC for the threshold method applied in this study. Moreover, a semiautomated approach would be less time- and labor-consuming, especially for research studies with a large sample size. The voxelwise method has the potential to explore the (sub)regional variation of the SN, yielding a substantial combined CR_V mean ICC, although some individual voxels, in particular those close to the borders of the ROI, showed somewhat lower reliability.

Limitations

In the current study all imaging slabs were placed by one person (M.P.) according to a commonly used method to reduce variation However, we still had to exclude one subject for the voxelwise analysis due to suboptimal placement. Correct placement of the NM-MRI is challenging and susceptible to differences in acquisition in and between studies. This would suggest that using a detailed volume placement protocol, such as that described recently,[33] is advisable to increase the reproducibility of the NM measurements. Furthermore, the current raters were inexperienced with NM-MRI segmentation. We tried to overcome this by employing a segmentation protocol and a training session. However, reliability might increase with more experience.

Another limitation might be that MRI scan parameters such as TR and TE differed between sequences and that adjustments might affect the reproducibility. The sequences that we applied were based on previously published sequences. Unpublished data from our lab has demonstrated that increasing the TR (to 633 msec) for the GRE MT on sequence leads to higher CR, which might improve reliability. This is in line with data by Wengler et al, who showed higher CR at a higher TR,[33] although in that study, other parameters such as slice thickness, were also adjusted simultaneously. In addition, a recent study by Liu et al showed that optimizing the FA increased the contrast-to-noise, which could also affect the reproducibility.[39] They found a different optimal FA for imaging the LC than for the SN. This means that optimization of NM-MRI for SN measurements may not be the same as for LC measurements. Since the LC is also an important structure that can be visualized with NM-MRI, separate (or simultaneous) studies assessing the reproducibility and optimization of NM-MRI of the LC are necessary.[31,34,40] Further optimization of the sequences by adjusting different MRI acquisition parameters

is therefore of interest. Combining such an optimization study with postmortem research is meaningful to evaluate the correlation with regional NM concentration and to evaluate the reliability for quantification of neuronal loss, for instance in PD.

Due to the inclusion of only relatively young and healthy subjects, the results of this study might not be generalizable to other populations. In addition, when using a clinical sample, such as PD patients, the reproducibility of the SN may be decreased, since the NM signal in PD patients is lower, leading to lower contrast and clinical symptoms may introduce (movement) artifacts in NM-MRI images.

Conclusion

NM-MRI CR is a highly reproducible measure, especially when using the GRE MT on sequence. Different analysis methods can be applied for CR analyses; however, for volume analyses the manual method is unreliable, whereas a thresholding method shows good results. Future research with the GRE MT on sequence is encouraged to further optimize NM-MRI as a noninvasive measurement of neuromelanin in the SN as a proxy biomarker for functioning of the dopamine system in different neuropathology.

REFERENCES

1. Sulzer D, Cassidy C, Horga G, et al. Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease. NPJ Park Dis 2018;4 11:
2. Shibata E, Sasaki M, Tohyama K, et al. Use of neuromelanin-sensitive MRI to distinguish schizophrenic and depressive patients and healthy individuals based on signal alterations in the substantia nigra and locus ceruleus. Biol Psychiatry 2008;64:401-406.
3. Lee J, Park S. Working memory impairments in schizophrenia: A meta-analysis. J Abnorm Psychol 2005;114:599-611.
4. Trujillo P, Summers PE, Ferrari E, et al. Contrast mechanisms associated with neuromelanin-MRI. Magn Reson Med 2017;78:1790-1800.
5. Matsuura K, Maeda M, Yata K, et al. Neuromelanin magnetic resonance imaging in Parkinson's disease and multiple system atrophy. Eur Neurol 2013;70:70-77.
6. Cassidy CM, Zucca FA, Girgis RR, et al. Neuromelanin-sensitive MRI as a noninvasive proxy measure of dopamine function in the human brain. Proc Natl Acad Sci U S A 2019;116:5108-5117.
7. Sulzer D, Bogulavsky J, Larsen KE, et al. Neuromelanin biosynthesis is driven by excess cytosolic catecholamines not accumulated by synaptic vesicles. Proc Natl Acad Sci U S A 2000;97:11869-11874.
8. Biesemeier A, Eibl O, Eswara S, et al. Elemental mapping of neuromelanin organelles of human substantia nigra: Correlative ultrastructural and chemical analysis by analytical transmission electron microscopy and nano-secondary ion mass spectrometry. J Neurochem 2016;138(2): 339-353.
9. Zucca FA, Basso E, Cupaioli FA, et al. Neuromelanin of the human substantia nigra: An update. Neurotox Res 2014;25:13-23.
10. Zecca L, Stroppolo A, Gatti A, et al. The role of iron and molecules in the neuronal vulnerability of locus coeruleus and substantia nigra during aging. Proc Natl Acad Sci U S A 2004;101:9843-9848.
11. Xing Y, Sapuan A, Dineen RA, Auer DP. Life span pigmentation changes of the substantia nigra detected by neuromelanin-sensitive MRI. Mov Disord 2018;33:1792-1799.
12. Moon WJ, Park JY, Yun WS, et al. A comparison of substantia nigra T1 hyperintensity in parkinson's disease dementia, Alzheimer's disease and age-matched controls: Volumetric analysis of neuromelanin imaging. Korean J Radiol 2016;17:633-640.
13. Hatano T, Okuzumi A, Kamagata K, et al. Neuromelanin MRI is useful for monitoring motor complications in Parkinson's and PARK2 disease. J Neural Transm 2017;124:407-415.
14. Kashihara K, Shinya T, Higaki F. Neuromelanin magnetic resonance imaging of nigral volume loss in patients with Parkinson's disease. J Clin Neurosci 2011;18:1093-1096.
15. Martín-Bastida A, Lao-Kaim NP, Roussakis AA, et al. Relationship between neuromelanin and dopamine terminals within the Parkinson's nigrostriatal system. Brain 2019;142:2023-2036.
16. Miyoshi F, Ogawa T, Kitao SI, et al. Evaluation of Parkinson disease and Alzheimer disease with the use of neuromelanin MR imaging and123I-metaiodobenzylguanidine scintigraphy. Am J Neuroradiol 2013;34:2113-2118.
17. Reimão S, Pita Lobo P, Neutel D, et al. Substantia nigra neuromelanin magnetic resonance imaging in de novo Parkinson's disease patients. Eur J Neurol 2015;22:540-546.
18. Isaias IU, Trujillo P, Summers P, et al. Neuromelanin imaging and dopaminergic loss in Parkinson's disease. Front Aging Neurosci 2016;8:1-12.
19. Kawaguchi H, Shimada H, Kodaka F, et al. Principal component analysis of multimodal neuromelanin MRI and dopamine transporter PET data provides a specific metric for the nigral dopaminergic neuronal density. PLoS One 2016;11:1-13.
20. Sasaki M, Shibata E, Tohyama K, et al. Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in Parkinson's disease. Neuroreport 2006;17:1215-1218.
21. Pyatigorskaya N, Gaurav R, Arnaldi D, et al. Magnetic resonance imaging biomarkers to assess substantia nigra damage in idiopathic rapid eye movement sleep behavior disorder. Sleep 2017;40 11:
22. Pyatigorskaya N, Magnin B, Mongin M, et al. Comparative study of MRI biomarkers in the substantia nigra to discriminate idiopathic Parkinson disease. Am J Neuroradiol 2018;39:1460-1467.
23. Schwarz ST, Xing Y, Tomar P, Bajaj N, Auer DP. In vivo assessment of brainstem depigmentation in Parkinson disease: Potential as a severity marker for multicenter studies. Radiology 2017;283:789-798.
24. Schwarz ST, Rittman T, Gontu V, Morgan PS, Bajaj N, Auer DP. T1-weighted MRI shows stage-dependent substantia nigra signal loss in Parkinson's disease. Mov Disord 2011;26:1633-1638.
25. Ariz M, Abad RC, Castellanos G, et al. Dynamic atlas-based segmentation and quantification of neuromelanin-rich brainstem structures in Parkinson disease. IEEE Trans Med Imaging 2019;38:813-823.
26. Fabbri M, Reimão S, Carvalho M, et al. Substantia nigra neuromelanin as an imaging biomarker of disease progression in Parkinson's disease. J Parkinsons Dis 2017;7:491-501.
27. Ohtsuka C, Sasaki M, Konno K, et al. Changes in substantia nigra and locus coeruleus in patients with early-stage Parkinson's disease using neuromelanin-sensitive MR imaging. Neurosci Lett 2013;541:93-98.
28. Langley J, Huddleston DE, Chen X, Sedlacik J, Zachariah N, Hu X. A multicontrast approach for comprehensive imaging of substantia nigra. Neuroimage 2015;112:7-13.
29. Huddleston DE, Langley J, Dusek P, et al. Imaging Parkinsonian pathology in substantia nigra with MRI. Curr Radiol Rep 2018;6:15. https://doi.org/10.1007/s40134-018-0272-x.
30. Huddleston DE, Langley J, Sedlacik J, Boelmans K, Factor SA, Hu XP. In vivo detection of lateral–ventral tier nigral degeneration in Parkinson's disease. Hum Brain Mapp 2017;38:2627-2634.
31. Chen X, Huddleston DE, Langley J, et al. Simultaneous imaging of locus coeruleus and substantia nigra with a quantitative neuromelanin MRI approach. Magn Reson Imaging 2014;32:1301-1306.
32. Langley J, Huddleston DE, Liu CJ, Hu X. Reproducibility of locus coeruleus and substantia nigra imaging with neuromelanin sensitive MRI. Magn Reson Mater Phys Biol Med 2017;30:121-125.

33. Wengler K, He X, Abi-Dargham A, Horga G. Reproducibility assessment of neuromelanin-sensitive magnetic resonance imaging protocols for region-of-interest and voxelwise analyses. Neuroimage 2019;208: 1-37.

34. Tona KD, Keuken MC, de Rover M, et al. In vivo visualization of the locus coeruleus in humans: Quantifying the test–retest reliability. Brain Struct Funct 2017;222:4203-4217.

35. Ohtsuka C, Sasaki M, Konno K, et al. Differentiation of early-stage parkinsonisms using neuromelanin-sensitive magnetic resonance imaging. Park Relat Disord 2014;20:755-760.

36. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31:1116-1128.

37. Landis R, Koch G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363-374.

38. Trujillo P, Smith AK, Summers PE, et al. High-resolution quantitative imaging of the substantia nigra. Proc Annu Int Conf IEEE Eng Med Biol Soc 2015;2015: 5428-5431.

39. Liu Y, Li J, He N, et al. Optimizing neuromelanin contrast in the substantia nigra and locus coeruleus using a magnetization transfer contrast prepared 3D gradient recalled echo sequence. Neuroimage 2020; 218:116935.

40. Betts MJ, Cardenas-Blanco A, Kanowski M, Jessen F, Düzel E. In vivo MRI assessment of the human locus coeruleus along its rostrocaudal extent in young and older adults. Neuroimage 2017;163:150-159.