Published October 4, 2024 as 10.3174/ajnr.A8526

ORIGINAL RESEARCH

Super-Resolution in Clinically Available Spinal Cord MRIs Enables Automated Atrophy Analysis

Blake E. Dewey, Samuel W. Remedios, Muraleetharan Sanjayan, Nicole Bou Rjeily, Alexandra Zambriczki Lee, Chelsea Wyche, Safiya Duncan, Jerry L. Prince, Peter. A. Calabresi, Kathryn C. Fitzgerald, Ellen M. Mowry

ABSTRACT

BACKGROUND AND PURPOSE: Measurement of the mean upper cervical cord area (MUCCA) is an important biomarker in the study of neurodegeneration. However, dedicated high-resolution scans of the cervical spinal cord are rare in standard-of-care imaging due to timing and clinical usability. Most clinical cervical spinal cord imaging is sagittally acquired in 2D with thick slices and anisotropic voxels. As a solution, previous work describes high-resolution T1-weighted brain imaging for measuring the upper cord area, but this is still not common in clinical care.

MATERIALS AND METHODS: We propose using a zero-shot super-resolution technique, SMORE, already validated in the brain, to enhance the resolution of 2D-acquired scans for upper cord area calculations. To incorporate super-resolution in spinal cord analysis, we validate SMORE against high-resolution research imaging and in a real-world longitudinal data analysis.

RESULTS: Super-resolved images reconstructed using SMORE showed significantly greater similarity to the ground truth than low-resolution images across all tested resolutions ($p<0.001$ for all resolutions in PSNR and MSSIM). MUCCA results from super-resolved scans demonstrate excellent correlation with high-resolution scans ($r>0.973$ for all resolutions) compared to low-resolution scans. Additionally, super-resolved scans are consistent between resolutions ($r>0.969$), an essential factor in longitudinal analysis. Compared to clinical outcomes such as walking speed or disease severity, MUCCA values from low-resolution scans have significantly lower correlations than those from high-resolution scans. Super-resolved results have no significant difference. In a longitudinal real-world dataset, we show that these super-resolved volumes can be used in conjunction with T1-weighted brain scans to show a significant rate of atrophy (-0.790, $p=0.020$ vs. -0.438, $p=0.301$ with low-resolution).

CONCLUSIONS: Super-resolution is a valuable tool for enabling large-scale studies of cord atrophy, as low-resolution images acquired in clinical practice are common and available.

ABBREVIATIONS: MS=multiple sclerosis; MUCCA=mean upper cervical cord; HR=high-resolution; LR=low-resolution; SR=super-resolved; CSC=cervical spinal cord; PMJ=pontomedullary junction; MSSIM=mean structural similarity; PSNR=peak signal-to-noise ratio; EDSS=expanded disability status scale.

Received month day, year; accepted after revision month day, year.
From the Department of Neurology (B.E.D., M.S., N.B., A.Z.L., C.W., S.D., P.A.C., K.C.F., E.M.M.), Department of Computer Science (S.W.R.) and Department of Electrical and Computer Engineering (J.L.P.) Johns Hopkins University, Baltimore, Maryland, U.S.A

Peter A. Calabresi has received personal consulting fees from Biogen, is a PI on grants to JHU from Annexon and Biogen, and consults for Disarm Therapeutics. Ellen M. Mowry receives research funding from Genentech and Biogen, consults for BeCareLink LLC, and receives royalties for editorial duties from UpToDate. All other authors declare no conflicts of interest related to the content of this article.

Please address correspondence to Blake E. Dewey, Ph.D., Department of Neurology, Johns Hopkins University, Pathology 627, 600 N. Wolfe Street, Baltimore, Maryland 21287, U.S.A.; blake.dewey@jhu.edu.

SUMMARY SECTION

PREVIOUS LITERATURE: The spinal cord is a critical target for investigation in MS. Previous works have described the calculation of spinal cord measurements from high-resolution spinal cord and brain images but have not yet explored clinically acquired spinal cord scans, which differ in resolution and acquisition. In the brain, super-resolution techniques, such as SMORE, have been shown to improve the reliability and accuracy of automated algorithms on images with low-resolution, isotropic voxels.

KEY FINDINGS: Super-resolution enables quantitative analysis of spinal cord MRI, even in cases of anisotropic voxels and slice gaps. Super-resolved images produce results on par with high-resolution results and can be used in analysis with high-resolution images of the brain and spinal cord for atrophy analysis..

KNOWLEDGE ADVANCEMENT: We have learned that super-resolution techniques can advance quantitative analysis for large-scale clinical studies. With this knowledge, previous limitations in image analysis can be questioned, and new, more extensive studies can be conducted with greater inclusivity and depth.

INTRODUCTION

Magnetic resonance imaging (MRI) is a commonly used imaging modality for diagnosis, monitoring, and prognostication in people living with neurodegenerative diseases such as multiple sclerosis (MS)[1–7]. While the bulk of imaging in clinical research has focused on the brain and its substructures, a growing community is investigating the spinal cord in the context of neurodegenerative diseases[8–12]. The mean upper cervical cord area (MUCCA) has been shown in the literature to be strongly correlated to disability, especially as related to motor

Copyright 2024 by American Society of Neuroradiology.

and sensory tasks[13–15]. However, wide dissemination of this measurement remains limited due to the extreme rarity of dedicated high-resolution (HR) spinal cord imaging in practice, where low-resolution (LR) clinical imaging or brain imaging predominates.

The whole spinal cord can be clearly delineated from the surrounding cerebrospinal fluid (CSF) using HR, isotropic T2-weighted (T2w) imaging with long echo times to minimize intra-cord contrast[16,17]. These scans are well standardized but take 4-5 minutes to acquire and have limited clinical utility, reducing their feasibility in clinical settings. In research settings, dedicated spinal cord scanning is still uncommon, with multiple groups proposing to use specific HR T1-weighted (T1w) brain scans already acquired in brain studies, including the upper part of the cervical spinal cord (CSC)[18,19]. These scans are common in research settings and are increasingly being adopted for clinical imaging. However, their implementation is still limited to research-centric clinical centers and needs widespread adoption[20]. Additionally, the acquisition (T1w vs. T2w) affects the results of automated spinal cord segmentation due to the different appearance of tissues like the CSF, dura matter, and white matter lesions and differences in image generation like partial voluming[19,21]. MUCCA measurements from T1w and T2w images are highly correlated, but adjustment would be required to use them interchangeably in longitudinal analyses.

Clinically, sagittal T2w images of the CSC are more common. However, these images are 2D-acquired with thick slices and sometimes a gap between the slices. For example, some of the highest resolution clinical spinal cord images are 3||0 (read "3-skip-0"), indicating a 3mm slice thickness and no gap (0mm), and are commonly acquired with the resolution 3||1 (3mm slice with 1mm gap). This kind of imaging is unsuitable for quantitative evaluation due to the measurement variation across resolutions and subjects.

Synthetic Multi-Orientation Resolution Enhancement (SMORE) is a self-supervised zero-shot super-resolution technique designed to enhance the resolution of anisotropic acquisitions[22,23]. SMORE has been extensively validated in the brain but has yet to be explored in the spinal cord. As SMORE is a zero-shot method, it requires no external training data. Instead, the training data are simulated from the target image, and training is performed on the simulation data (hence, self-supervised). This means that SMORE can be applied to a new image contrast or body part without collecting training data or worrying about training/testing bias. This differs from other super-resolution approaches (like SynthSR[24], TSCTNet[25], and others[26]), which utilize extensive training datasets and are currently focused on brain imaging.

This work aims to demonstrate super-resolution as a tool to enable MUCCA estimation on clinically available LR spinal cord images. We make two important contributions:

1. Demonstrate improved outcomes when using super-resolved images for MUCCA calculation compared to LR images using simulated datasets with HR ground truth.
2. Measure CSC atrophy in a real-world longitudinal dataset with super-resolved LR 2D spine and HR 3D brain images.

The results from this work set the stage for large-scale studies of CSC atrophy, which can be conducted at reduced cost and with increased availability by using existing clinically acquired imaging datasets.

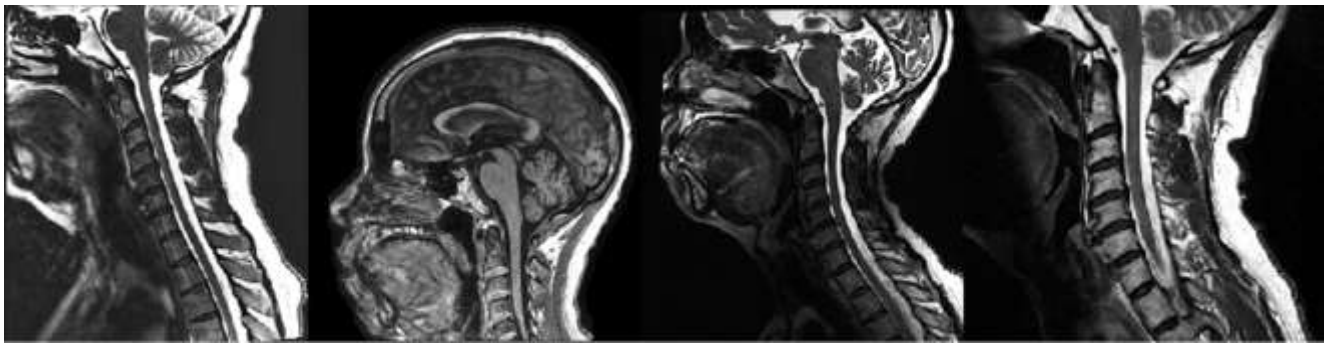

FIG 1. A representative history of one subject from the Real-World Longitudinal Dataset. The four images (left to right): 2D LR T2w CSC (3||0.5), 3D T1w Brain, 3D T2w CSC, 2D LR T2w CSC (3||0).

MATERIALS AND METHODS

Imaging Datasets

HR Research Dataset

Paired LR and HR images are rarely acquired, especially in the CSC. To validate super-resolution techniques quantitatively, we simulate LR data from acquired HR data. To this end, we selected 200 participants who underwent a research MRI protocol on a single Siemens Prisma scanner as a part of an existing Institutional Review Board-approved study of people with MS. Imaging included T1w 3D Magnetization Prepared Rapid Gradient Echo (MPRAGE) of the brain [resolution: 1mm isotropic, orientation: sagittal, field of view (FOV): 256x240x160mm, echo time (TE): 2.98ms, repetition time (TR): 2300ms, inversion time (TI): 900ms, flip angle (FA): 9deg, acceleration: 2, acquisition time (AT): 5:12 min) and T2w 3D Turbo Spin Echo (T2-SPACE) of the CSC [resolution: 0.8mm isotropic, orientation: sagittal, FOV: 256x256x64mm, TE: 120ms, TR: 1500ms, FA: 120deg, averages: 1.4, acceleration: 3, AT: 4:02 min]. These images comprised our "HR Research Dataset" and are high-resolution isotropic volumes, which allow us to create simulated LR images and still provide a ground truth for quantitative assessment. Additionally, the paired brain and CSC images will enable us to quantitatively compare results from HR 3D T1w brain and HR 3D T2w CSC images.

Real-World Longitudinal Dataset

Simulated LR data is insufficient to evaluate super-resolution's effect in a real-world longitudinal study. To validate our methodology in a real-world example, we created a sub-cohort (N=130) from PwMS in the IRB approved study that had multiple available clinical brain and/or spinal cord MRIs acquired between 2013 and 2023. Using the Johns Hopkins Precision Medicine Access Platform, we retrieved all brain and spinal cord scan sessions for each sub-cohort participant from the clinical imaging system collected over these ten years. Each scan session contributed one volume to the analysis: a 3D T1w MPRAGE brain image, a 3D T2w CSC image, or a 2D LR T2w CSC image. Each of these images was directly acquired on a clinical scanner. This "Real-World Longitudinal Dataset" consisted of 700 images with an average of 5.6 images per person and an average follow-up of 4.1 years. In terms of image acquisition, 315 (45%) images were LR 2D T2w CSC, 180 (26%) images were HR 3D T2w CSC, and 205 (29%) images were HR 3D T1w brain. Representative images from the Real-World Longitudinal Dataset are shown in Figure 1.

Blinded Clinical Testing

All participants underwent the MS Functional Composite (MSFC), which is composed of three separate tasks: a timed 25-foot walk (T25FW), a 9-hole peg test (9HPT), and a paced auditory serial addition test [27]. For this analysis, we focused on the motor-associated tasks T25FTW and 9HPT, hypothesized to be the most relevant to spinal cord atrophy. Additionally, each participant was scored using the Expanded Disability Status Scale (EDSS)[28], which favors motor disability in its scoring.

Cord Segmentation and MUCCA Calculation

The Spinal Cord Toolbox[29] (v6.0) was used for all spinal cord segmentations and analyses. The spinal cord was segmented using SCT's DeepSeg[30] algorithm. Then, MUCCA was calculated by averaging the cross-sectional area over 3cm beginning 6cm below the pontomedullary junction (PMJ) as described by Bédard et al. [31]. This was empirically more stable than averaging over the C2-C3 levels, as it did not require segmentation of the vertebral levels, and vertebral segmentation often required manual intervention, especially in T1w brain images. All volumes, including T1w brain and LR T2w CSC acquisitions, were segmented using this method. Quality assurance was done manually by a single rater (B.D.) to ensure high-quality segmentation. In <5% of cases, the PMJ had to be manually delineated. We selected manual PMJ landmarks using a graphical viewer in SCT, which took one rater (B.D.) less than 15 minutes for all missed cases. Example segmentations are shown in Figure 2.

Super-Resolution

Super-resolution with SMORE has two main steps: training and inference. As SMORE is a zero-shot, internally trained method, it must be trained on each image. SMORE is designed for super-resolution on anisotropic images, meaning that the resolution of the 3D volumes has two high-resolution "in-plane" directions and one low-resolution "through-plane" (or slice) direction. Training in SMORE takes advantage of this fact by degrading the high-resolution in-plane slices in one direction to simulate the appearance of a through-plane slice. Simulated low-resolution patches are generated using the `degrade` feature of the `radifox-utils` Python package (https://github.com/jh-mipc/radifox-utils) to apply a learned slice profile to the high-resolution patches. This relative slice profile is predicted using Estimating the Slice Profile for Resolution Enhancement of a Single image Only (ESPRESO)[32], which uses adversarial learning to produce a slice profile that generates similar distributions of real and simulated through-plane patches. After degradation, these simulated low-resolution and real high-resolution pairs train a convolutional neural network to generate high-resolution patches. Once the model is trained, real through-plane slices are passed through the network to generate the super-resolved slices. SMORE was implemented using v4.0.5 of the open-source software (https://gitlab.com/iacl/smore).

SMORE Validation

To generate validation data from the HR Research Dataset, the HR T2-SPACE images were artificially degraded to match the four most common resolutions found in our clinical system for sagittal CSC images: 3||0, 3||0.3, 3||0.5, 3||1. Degradation was performed using the `degrade` function of the `radifox-utils` package. This blurred the image using a real-world slice profile constructed with the Shinnar–Le Roux algorithm[33] according to the slice thickness, then downsampled the image according to the slice spacing. This is a more accurate simulation of a 2D-acquired image than downsampling alone because it more closely approximates the actual acquisition process of an MRI.

Super-resolved (SR) and LR images were compared to HR ground truth images using Mean Structural Similarity (MSSIM)[34] and Peak Signal-to-Noise Ratio (PSNR)[35]. LR images were interpolated to the HR grid using the `resize` function of `radifox-utils` and a 3rd-order B-Spline for these comparisons. Each SR, LR, and HR image was also segmented, and MUCCA was calculated. MUCCA measurements from SR and LR images were compared to the HR results. Pearson's rho was used to determine the correlation between SR (or LR) and HR results at each simulated resolution. A paired Student's t-test was used to determine whether the differences between SR/LR and HR MUCCA values and between image metrics (PSNR and MSSIM) were statistically significant.

To determine the effect of super-resolution on outcomes in a clinical study, MUCCA values from LR, SR, and HR images were modeled as predictors of clinical outcomes using linear regression models and partial correlation using Pearson's method. Willam's test was used to determine the significance of correlation differences. Simulated real-world cohorts were created from the LR and SR datasets by randomly selecting a resolution (3||0.0, 3||0.3, 3||0.5, or 3||1.0) for each participant. This simulates a real-world dataset that might contain acquisitions acquired at different resolutions. Models and correlations were adjusted for age and sex at birth. The size of our dataset could result in significant results that might not hold up in smaller samples. To evaluate this, we created 100 bootstrapped samples of 50 subjects to evaluate the effect of sample size on the significance of the relationships.

Longitudinal Analysis

<u>Correction for T1w Brain Results</u>

The HR T1w brain scans in the HR Research Dataset were segmented, and MUCCA was calculated for each image. A linear fit of MUCCA from T2w spinal cord images versus MUCCA from T1w brain images was used to determine an additive correction factor. This corrective factor was applied to all T1w values in the longitudinal cohort.

Modeling Atrophy

Each 2D T2-weighted cervical spinal cord image from the Real-World Longitudinal Dataset was super-resolved using SMORE. Then, all super-resolved 2D and acquired 3D images (brain or spine) in that dataset were segmented, and MUCCA was calculated for each image. A linear mixed effects model of MUCCA vs. time was fit, adjusting for age at first scan and sex.

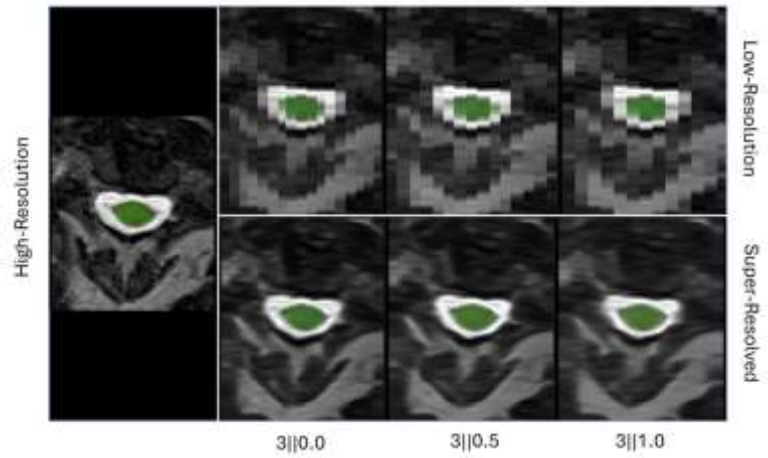

FIG 2. Segmentation of representative HR, simulated LR, and SR volumes.

RESULTS

Qualitative Evaluation

As shown in Figure 2, SMORE substantially recovers the spinal cord's anatomical structure. This is reflected in the segmentation quality, as SCT is not only limited to the appearance of the spinal cord in the image (after interpolation within the algorithm) but also in the final resolution of the output. LR inputs produce blocky segmentations that match the image resolution. We found that interpolation before segmentation with SCT to avoid this difference made results substantially worse, likely due to the additional internal interpolation step within SCT. In the SCT segmentations, we noted four subjects where some LR images had poor-quality segmentation. In contrast, all SR and HR images were correctly segmented. As stated above, some images required manual delineation of the PMJ; this was mainly due to anatomical variation in the subjects and was the same in all images regardless of resolution or preparation. However, there were a few LR images at 3||1 where the PMJ was not correctly delineated when it was correct on the HR image.

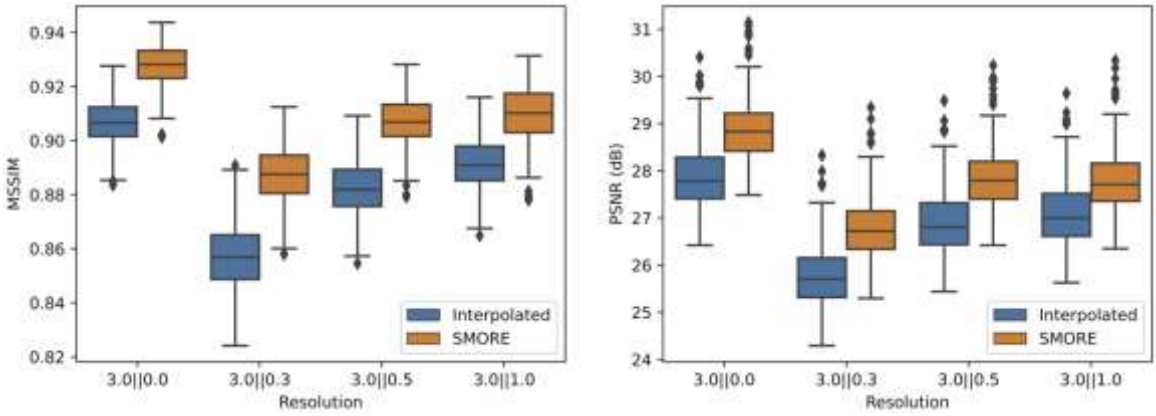

FIG 3. Boxplots of MSSIM and PSNR values calculated for interpolated LR and SR images at each resolution compared to HR ground truth.

Quantitative Validation of SMORE

As previously shown in the brain, SR images are more similar to the ground truth than degraded images by MSSIM and PSNR (Figure 3). This difference is statistically significant across all tested resolutions. However, downstream analysis feasibility depends more on the quality of the segmentation results than image quality.

Figure 4 shows that low-resolution and super-resolved results have an excellent correlation with high-resolution results. The super-resolved correlation results are nearly 1, even in the 3||1 case, while the low-resolution results drop to 0.92. The super-resolved results also

demonstrate an excellent correlation between super-resolved images of different resolutions. While low-resolution results maintain some areas of excellent correlation between resolutions, some pairings begin to dip below 0.9.

In Figure 5, we directly plot the relationship between SR/LR results and the corresponding HR results, including an identity line. We can see that the SR values are tightly following and not significantly different from the identity line (3||0.0: p=0.12, R^2=0.98; 3||0.3: p=0.21, R^2=0.95; 3||0.5: p=0.77, R^2=0.96; 3||1.0: p=0.15, R^2=0.96), with some increased variation around that as the slice gap increases to 1mm. The LR results show over-segmentation of the cord compared to HR, which increases with cord area. This is evident across all resolutions and is most extreme in 3||0.3 and 3||1.0 cases. This deviation in slopes from identity is also statistically significant (p<0.0001 for all resolutions, R^2=0.91, 0.90, 0.86, 0.85 for each resolution, respectively).

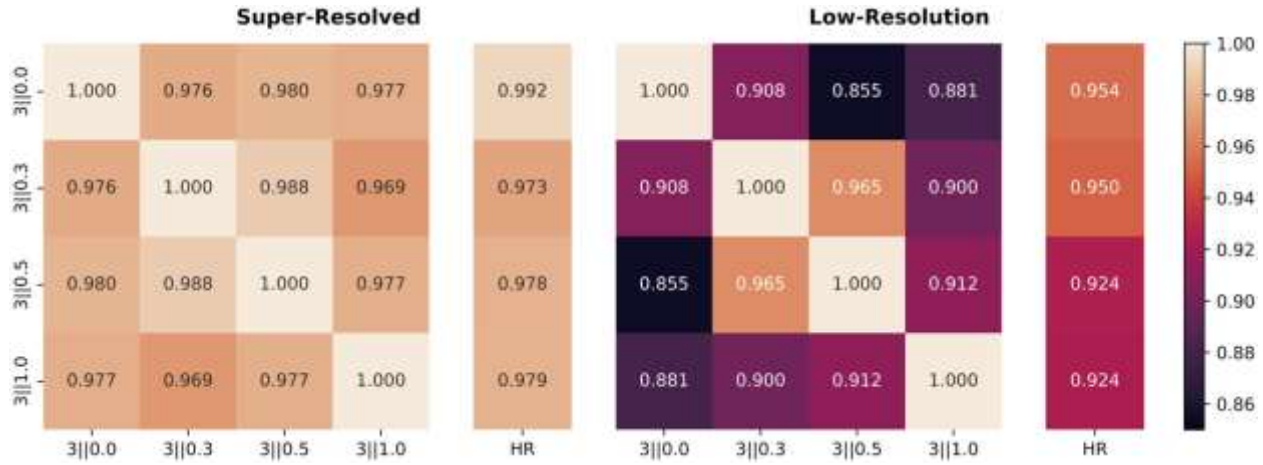

FIG 4. Heat maps showing the correlation between each simulated resolution and other simulated resolutions, as well as with ground-truth high-resolution for both super-resolved and low-resolution images.

Clinical Correlations

Figure 6 shows the correlations between MUCCA values derived from the HR, LR, and SR randomized resolution datasets and clinical outcomes. For each of these outcomes, correlations to MUCCA values are statistically significant (EDSS: LR r=-0.22/p=0.010, SR r=-0.30/p=<0.001, HR r=-0.30/p=<0.001; 25FTW: LR r=-0.22/p=0.014, SR r=-0.26/p=0.003, HR r=-0.27/p=0.002; 9HPT: LR r=-0.29/p=<0.001, SR r=-0.38/p=<0.001, HR r=-0.38/p=<0.001). Additionally, LR correlations were significantly less than the HR correlations (EDSS: p=0.04, 25FTW: p=0.046, 9HPT: p<0.001), while the SR correlations were not significantly different. In linear modeling, the slope of the effect from LR MUCCA values is underestimated compared to HR values (EDSS: LR=-0.033, HR=-0.052; 25FTW: LR=-0.009, HR=-0.014; 9HPT: LR=-0.250, HR=-0.394). However, the effect slope for SR values is nearly identical to the effect of HR values (EDSS: SR=-0.051, 25FTW: SR=-0.013, 9HPT=SR=-0.388). Additionally, the significance of the LR relationships did not hold up in smaller samples. For example, our bootstrapped samples yielded an average p-value of 0.12 when comparing LR MUCCA values to 25FTW. At the same time, the HR and SR MUCCA values retain a significant relationship with 25FTW (average p=0.005).

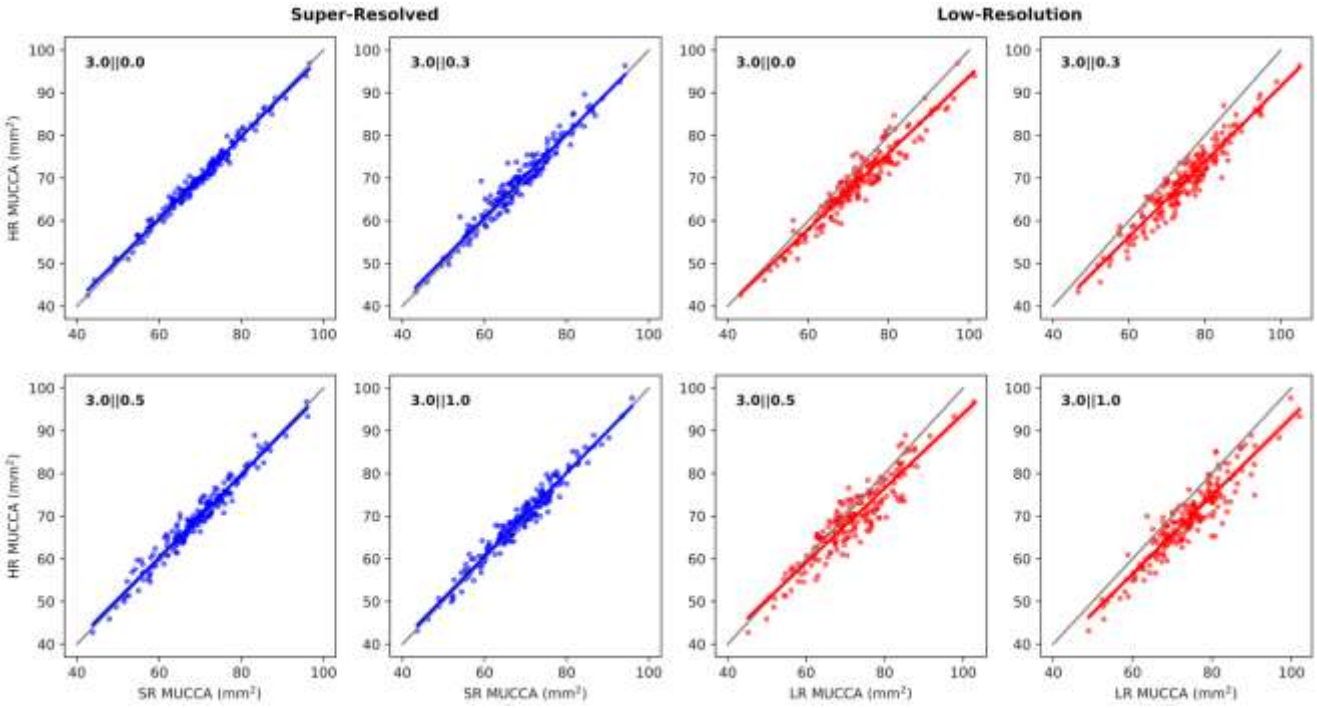

FIG 5. Scatter plots of super-resolved (left) and low-resolution (right) vs. high-resolution MUCCA values. The line of fit and 95%

confidence intervals are plotted in the corresponding color, and the identity line is plotted in grey.

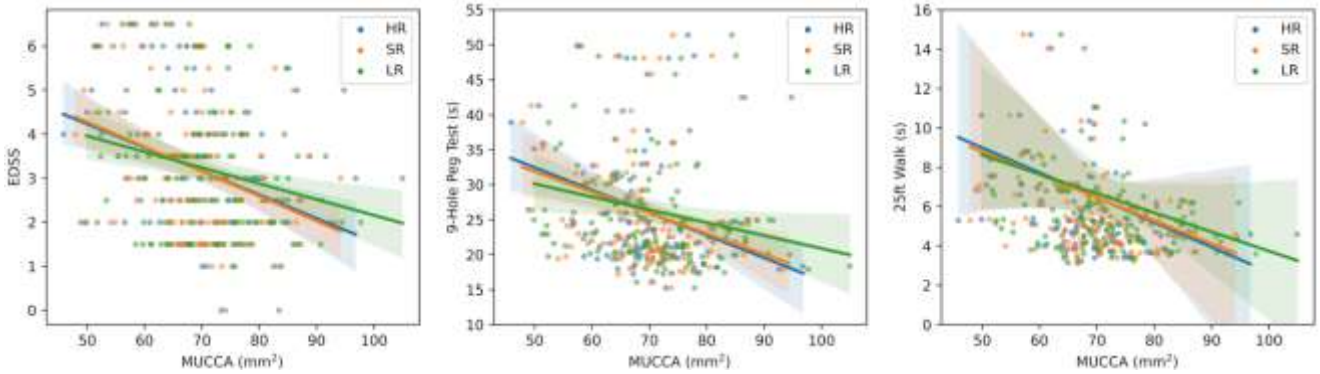

FIG 6. Scatter plots of MUCCA values versus clinical outcomes: EDSS (left), 9-hole peg test (middle), and 25-ft walk (right). Each plot shows points, lines of fit, and 95% confidence intervals for HR (blue), SR (orange), and LR (green).

T1w Brain Comparison

Figure 7 shows the relationship between MUCCA results from T1w brain images and T2w CSC images from the HR Research Dataset. These measures are strongly correlated (ρ=0.974). However, there is also a clear and substantial bias (β=9.184). We can adjust for this bias by adding a correction factor to all T1w brain results. This allows the recovery of a near-identity relationship (slope not significantly different from 1, p=0.100). This adjustment was used to correct the T1w brain results in the longitudinal analysis.

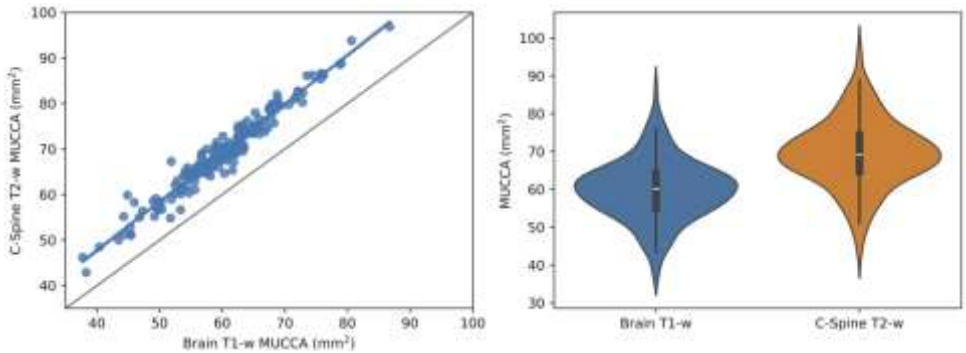

FIG 7. Left: Scatter plot of HR brain vs. HR spine MUCCA value, Right: Violin plot showing the distribution of MUCCA values from brain and spine datasets.

Longitudinal Analysis

Table 1 shows the fixed effects from the linear mixed effects model fit on the LR and SR versions of the Real-World Longitudinal Dataset. The decline of MUCCA over time (the "Time (from first scan)" effect) is lower than expected and not significant in either the LR or SR models. After inspecting the data, it was clear that the difference in contrast between the 2D-acquired and 3D-acquired images created a bias in the results. After adjusting for this bias (Table 2), the SR model now has a much larger time effect that is statistically significant. The cohort using low-resolution images also shows a slight increase in the time effect, but it is not significant. There is also a significant effect of age at first scan, sex, and age at first scan × time in the SR model.

Table 1: Results from linear mixed-effects models of MUCCA over time. "Low-Resolution" and "Super-Resolved" are separate models fit using the super-resolved or low-resolution versions of the longitudinal cohort. The "Time (from first scan)" fixed effect represents the change in MUCCA each year.

	Low-Resolution		Super-Resolved	
	coefficient	p-value	coefficient	p-value
Sex (M)	3.293	**0.040**	3.101	**0.048**
Age (at first scan, years)	-0.162	0.062	-0.122	0.154
Time (from first scan, years)	-0.356	0.360	-0.522	0.167
Age (at first scan) x Time (from first scan, years)	0.007	0.393	0.018	**0.020**

DISCUSSION

This paper demonstrates the ability of SMORE to super-resolve clinically available CSC MRIs, enabling reliable MUCCA calculation. This was validated directly in simulated experiments of the HR Research Dataset and the Real-World Longitudinal Dataset.

In simulated experiments, SMORE was able to recover qualitative anatomical features and improve quantitative similarity to HR ground truth images, especially around the spinal cord. We see similar improvement quantitatively, although all values are lower than in previous

work with SMORE in the brain[23], indicating a greater effect of differences in resolution on the degraded anatomy. The spinal cord is a small structure in a large cervical spine MRI, so structures outside the spinal canal can likely explain differences in qualitative and quantitative results.

Table 2: . Results from linear mixed-effects models of MUCCA over time, including a correction for 2D spinal cord images. "Low-Resolution" and "Super-Resolved" are separate models fit using the super-resolved or low-resolution versions of the longitudinal cohort. The "Time (from first scan)" fixed effect represents the change in MUCCA each year.

	Low-Resolution		Super-Resolved	
	coefficient	p-value	coefficient	p-value
Sex (M)	3.253	**0.042**	3.180	**0.042**
Image Type (2D, LR or SR)	0.829	0.190	-4.577	**<0.001**
Age (at first scan, years)	-0.152	0.080	-0.174	**0.041**
Time (from first scan, years)	-0.438	0.301	-0.790	**0.020**
Age (at first scan) x Time (from first scan, years)	0.007	0.394	0.016	**0.029**

Acquisition parameters in clinical cohorts are highly variable. For this reason, interoperability between contrasts and resolutions is critical to successful longitudinal analysis of clinically acquired images. In a real-world clinical environment, it is common for resolution to change over time as patients get imaging at other locations or protocols are updated. Correlation between different resolutions is vital to the feasibility of this method over longitudinal follow-up. Even the LR images produce MUCCA values with high to excellent correlations to the HR ground truth and each other. Yet, they still underestimate correlations when compared to clinical outcomes. In these simulated experiments, near identity is required to maintain the clinical correlations with sufficient statistical power. In our exploration of clinical correlations, we also see the possibility of producing a statistically significant result that underestimates the true effect according to HR data. This large cohort (N=200) produced significant results with very low p-values; however, similar findings were not obtained in smaller random subsets.

In addition to SR clinical spine images, we include HR 3D T1w brain images in our definition of "clinically available MRI." While this is still uncommon in many clinical settings, using these images without contrast is becoming more popular, especially as reimbursement of quantitative image analysis is now possible for U.S. payers[36]. This also allows the frequent follow-up of MUCCA when brain MRIs are performed without spine imaging. In clinical datasets, different providers have different ordering preferences that can depend on the individual patient, so following patients with every scan possible enriches the available data pool.

Limitations

Analysis of our longitudinal cohort showed that super-resolved images play an important role in reliably quantifying atrophy. However, this analysis has limitations. As demonstrated by the statistical bias between 2D and 3D T2w images, there is a need to control for the differences in image contrasts. In this analysis, we performed statistical correction by adjusting for 2D image contrast in our mixed effects model, but other methods should also be explored to control for these differences. We also have no well-controlled validation for the effect sizes presented here. We plan to conduct this validation by collecting longitudinal HR spinal cord scans of research participants and comparing the results to longitudinal follow-ups that include other scan types. Additionally, we have not studied the effect of spinal cord lesions on the analysis. Lesion evolution is essential to MS pathology, especially in the spinal cord. In this work, we focused on the effects of atrophy, using T2w images where lesions are less apparent. However, these are inconsistent across acquisitions and may contribute to the differences in volumes in LR CSC images and the T1w brain images, where lesions could be mistaken for CSF more frequently than in the heavily T2w 3D CSC images.

Future Directions

We look forward to expanding this cohort in size and follow-up duration to investigate these findings further. We also look to collect more detailed clinical data over the participants' histories to correlate longitudinal patterns to clinical outcomes. In particular, we look to expand this analysis to include lesions to study the inflammatory pieces of the MS disease course. Ultimately, we aim to investigate MUCCA as a predictor and monitor of clinical progression.

CONCLUSIONS

In conclusion, we demonstrated the feasibility of MUCCA calculations after super-resolution from clinically available MRIs such as 2D-acquired T2w spinal cord images and 3D T1w brain images. We showed that SMORE produced super-resolved image volumes from 2D-acquired spinal cord scans with MUCCA values nearly identical to HR ground truth images. We also demonstrated that these, along with corrected values from T1w brain scans, can be used in a longitudinal analysis of spinal cord atrophy in people with MS. This opens the door to large, inclusive, clinically derived datasets for large-scale analysis of spinal cord atrophy.

ACKNOWLEDGMENTS

We want to thank the participants in this study, along with their families and caregivers, for their time and effort.

Funding

This work is funded by the National MS Society (FG-2008-36966 PI: Dewey, TA-1805-31136 PI: Fitzgerald), the National Institutes of Health (R01NR018851 PI: Mowry, R01NS082347 PI: Calabresi, K01MH121582 PI: Fitzgerald) and the National Science Foundation (DGE-1746891 PI: Remedios).

REFERENCES

1. Wattjes, M. P., Steenwijk, M. D. & Stangel, M. MRI in the Diagnosis and Monitoring of Multiple Sclerosis: An Update. *Clin. Neuroradiol.* **25**, 157–165 (2015).
2. Pretorius, P. M. & Quaghebeur, G. The Role of MRI in the Diagnosis of MS. *Clin. Radiol.* **58**, 434–448 (2003).
3. Wattjes, M. P. *et al.* 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol.* **20**, 653–670 (2021).
4. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018).
5. Cortese, R., Giorgio, A., Severa, G. & De Stefano, N. MRI Prognostic Factors in Multiple Sclerosis, Neuromyelitis Optica Spectrum Disorder, and Myelin Oligodendrocyte Antibody Disease. *Front. Neurol.* **12**, (2021).
6. Kaunzner, U. W. & Gauthier, S. A. MRI in the assessment and monitoring of multiple sclerosis: an update on best practice. *Ther. Adv. Neurol. Disord.* **10**, 247–261 (2017).
7. Hemond, C. C. & Bakshi, R. Magnetic Resonance Imaging in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **8**, (2018).
8. Chen, Y., Haacke, E. M. & Bernitsas, E. Imaging of the Spinal Cord in Multiple Sclerosis: Past, Present, Future. *Brain Sci.* **10**, 857 (2020).
9. Cohen, A. B. *et al.* The Relationships among MRI-Defined Spinal Cord Involvement, Brain Involvement, and Disability in Multiple Sclerosis. *J. Neuroimaging* **22**, 122–128 (2012).
10. Zeydan, B. *et al.* Cervical spinal cord atrophy. *Neurol. Neuroimmunol. Neuroinflammation* **5**, e435 (2018).
11. Bischof, A. *et al.* Spinal Cord Atrophy Predicts Progressive Disease in Relapsing Multiple Sclerosis. *Ann. Neurol.* **91**, 268–281 (2022).
12. Muccilli, A., Seyman, E. & Oh, J. Spinal Cord MRI in Multiple Sclerosis. *Neurol. Clin.* **36**, 35–57 (2018).
13. Daams, M. *et al.* Mean upper cervical cord area (MUCCA) measurement in long-standing multiple sclerosis: Relation to brain findings and clinical disability. *Mult. Scler. J.* **20**, 1860–1865 (2014).
14. Weeda, M. M. *et al.* Validation of mean upper cervical cord area (MUCCA) measurement techniques in multiple sclerosis (MS): High reproducibility and robustness to lesions, but large software and scanner effects. *NeuroImage Clin.* **24**, 101962 (2019).
15. Chien, C., Juenger, V., Scheel, M., Brandt, A. U. & Paul, F. Considerations for Mean Upper Cervical Cord Area Implementation in a Longitudinal MRI Setting: Methods, Interrater Reliability, and MRI Quality Control. *Am. J. Neuroradiol.* **41**, 343–350 (2020).
16. Cohen-Adad, J. *et al.* Generic acquisition protocol for quantitative MRI of the spinal cord. *Nat. Protoc.* **16**, 4611–4632 (2021).
17. Valošek, J. & Cohen-Adad, J. Reproducible Spinal Cord Quantitative MRI Analysis with the Spinal Cord Toolbox. *Magn. Reson. Med. Sci.* rev.2023-0159 (2024) doi:10.2463/mrms.rev.2023-0159.
18. Liu, Y. *et al.* Multicenter Validation of Mean Upper Cervical Cord Area Measurements from Head 3D T1-Weighted MR Imaging in Patients with Multiple Sclerosis. *Am. J. Neuroradiol.* **37**, 749–754 (2016).
19. Taheri, K. *et al.* Cervical Spinal Cord Atrophy can be Accurately Quantified Using Head Images. *Mult. Scler. J. - Exp. Transl. Clin.* **8**, 20552173211070760 (2022).
20. Saslow, L. *et al.* An International Standardized Magnetic Resonance Imaging Protocol for Diagnosis and Follow-up of Patients with Multiple Sclerosis: Advocacy, Dissemination, and Implementation Strategies. *Int. J. MS Care* **22**, 226–232 (2020).
21. Nagel, S. J. *et al.* Spinal dura mater: biophysical characteristics relevant to medical device development. *J. Med. Eng. Technol.* **42**, 128–139 (2018).
22. Zhao, C. *et al.* SMORE: A Self-supervised Anti-aliasing and Super-resolution Algorithm for MRI Using Deep Learning. *IEEE Trans. Med. Imaging* **40**, 805–817 (2021).
23. Remedios, S. W. *et al.* Self-Supervised Super-Resolution for Anisotropic MR Images with and Without Slice Gap. in *Simulation and Synthesis in Medical Imaging* (eds. Wolterink, J. M., Svoboda, D., Zhao, C. & Fernandez, V.) 118–128 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-44689-4_12.
24. Iglesias, J. E. *et al.* SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Sci. Adv.* **9**, eadd3607 (2023).
25. Lu, Z. *et al.* Two-Stage Self-Supervised Cycle-Consistency Transformer Network for Reducing Slice Gap in MR Images. *IEEE J. Biomed. Health Inform.* **27**, 3337–3348 (2023).
26. Du, J. *et al.* Super-resolution reconstruction of single anisotropic 3D MR images using residual convolutional neural network. *Neurocomputing* **392**, 209–220 (2020).
27. Fischer, J. S., Rudick, R. A., Cutter, G. R. & Reingold, S. C. The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. *Mult. Scler. J.* **5**, 244–250 (1999).
28. Kurtzke, J. F. Rating neurologic impairment in multiple sclerosis. *Neurology* **33**, 1444–1444 (1983).
29. De Leener, B. *et al.* SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *NeuroImage* **145**, 24–43 (2017).
30. Gros, C. *et al.* Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *NeuroImage Orlando Fla* **184**, 901–915 (2019).
31. Bédard, S. & Cohen-Adad, J. Automatic measure and normalization of spinal cord cross-sectional area using the pontomedullary junction. *Front. Neuroimaging* **1**, (2022).
32. Han, S., Remedios, S. W., Schär, M., Carass, A. & Prince, J. L. ESPRESO: An algorithm to estimate the slice profile of a single magnetic resonance image. *Magn. Reson. Imaging* **98**, 155–163 (2023).
33. Pauly, J., Le Roux, P., Nishimura, D. & Macovski, A. Parameter relations for the Shinnar-Le Roux selective excitation pulse design algorithm (NMR imaging). *IEEE Trans. Med. Imaging* **10**, 53–65 (1991).
34. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
35. Wang, Z. & Bovik, A. C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* **26**, 98–117 (2009).
36. Bash, S., Tanenbaum, L. N., Segovis, C. & Chen, M. CPT Codes for Quantitative MRI of the Brain: What It Means for Neuroradiology. *Am. J. Neuroradiol.* (2024) doi:10.3174/ajnr.A8286.

SUPPLEMENTAL FILES

TRIPOD+AI

Version: 11 January 2024

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	N/A
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	1
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	3
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	2
	3c	D;E	Describe any known health inequalities between sociodemographic groups	N/A
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	4
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	4
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	N/A
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	N/A
	6b	D;E	Describe the eligibility criteria for study participants	N/A
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	N/A
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	6
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	N/A
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	N/A
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	N/A
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	N/A
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	N/A
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	N/A
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	4
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	N/A
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	4
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	N/A
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	N/A
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	11
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	8-10
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	N/A
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	5
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	N/A
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	N/A
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	6

[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

[2] Separately for all model building approaches.

[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	4
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	N/A
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	13
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	2
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	N/A
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	N/A
Data sharing	18e	D;E	Provide details of the availability of the study data	N/A
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	6
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	N/A
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	N/A
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	N/A
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	N/A
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	N/A
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	6
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	8-10
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	N/A
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	N/A
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	10-11
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	11-12
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	11
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	N/A
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	12

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

[4] This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.
[5] This relates to the code to implement the model to get estimates of risk for a new individual.

N/A