VALIDATION OF QYSCORE® WHITE MATTER HYPERINTENSITY (WMH) U-NET SEGMENTATION ALGORITHM AGAINST EXPERT MANUAL CONSENSUS AND COMPARISON WITH STATE-OF-THE-ART METHODS

Elizabeth Gordon[1], Philippe Tran[1], Luca M. Villa[1], Mathilde Borrot[1], Ayoub Gueddou[1], Nicolas Guizard[1].

[1]Qynapse, 2-10 rue d'Oradour-sur-Glane, 75015 Paris, France

Email: egordon@qynapse.com

ALZHEIMER'S ASSOCIATION AAIC 24 Alzheimer's Association International Conference

QYNAPSE

BACKGROUND

Detection and quantification of White Matter Hyperintensities (WMH) are clinically important across multiple CNS disorders and neurodegenerative dementias. However, the labor-intensive nature of manual segmentation limits widespread clinical application. Validation of accurate automated methods for segmenting WMH are urgently needed to overcome this unmet clinical need.

OBJECTIVES

To validate QyScore®'s fully-automated WMH quantification pipeline against ground-truth expert manual consensus gold-standard, and directly compare performance accuracy against six widely used packages.

METHODS

The validation cohort consisted of 129 individuals who had undergone T1-weighted and T2-FLAIR MR imaging.

- To ensure robust results, *different scanners* (30 GE, 26 Philips, 73 Siemens) and *patient populations* were included (Table 1A).

The WMH_U-Net algorithm included in QyScore®, an FDA-cleared and CE-marked neuroimaging platform, automatically segmented WMH in each image set, using a convolutional neural network approach.

These were compared to the gold-standard consensus of three expert neuroradiologist manual segmentations to derive key performance metrics:

- spatial overlap (Dice Similarity Coefficient (**DSC**) and **F1** scores) and volume comparisons (intra-class correlation coefficient (**ICC**) and absolute volume error (**AVE**, ml).

A second investigation performed a *direct comparison* of QyScore® WMH_U-Net with six state-of-the-art supervised and unsupervised segmentation methods (*LST-LGA* and *LPA*, *Lesion-TOADS, lesionBrain, BIANCA* and *nicMSlesions*) on a dedicated MS dataset (Table 1B) with default and optimized settings where available. DSC, F1, ICC and AVE were compared across all methods.

Table 1A – Full validation cohort (n=129) demographics, split by lesion load and associated DSC and AVE performance metrics for the validation of QyScore® WMH_U-NET automated WMH segmentation algorithm against the consensus of three expert neuroradiologist manual segmentation

Algorithm	Lesion Load (based on expert manual consensus)	N of subjects	Age mean (std) [range]	Clinical Status	Sex (M – F)	Type 2D – 3D	DSC Results mean (std)	AVE Results mean (std)
QyScore® WMH_U-Net	WMH Low < 5 mL	29	58.13 (17.93) [26 – 90]	10 AD, 12 MS, 7 HC	12 – 17	18 – 11	0.36 (0.17)	0.76 (0.68)
QyScore® WMH_U-Net	WMH Medium 5 – 15 mL	23	59.85 (18.97) [29 – 84]	9 AD, 12 MS, 2 HC	11 – 12	14 – 9	0.68 (0.10)	3.29 (2.94)
QyScore® WMH_U-Net	WMH High 15 – 30 mL	46	63.41 (19.57) [27 – 91]	17 AD, 5 FTD, 13 MS, 5 HC, 6 N/A	24 – 22	23 – 23	0.75 (0.08)	4.38 (3.92)
QyScore® WMH_U-Net	WMH Very high > 30 mL	31	76.39 (12.13) [39 – 91]	13 AD, 8 MS, 6 HC, 4 NA	13 – 18	15 – 16	0.79 (0.05)	7.99 (5.46)
QyScore® WMH_U-Net	WMH **Full sample**	129	62.95 (19.35) [27 – 91]	49 AD, 5 FTD, 45 MS, 20 HC, 10 N/A	60 – 69	70 – 59	0.66 (0.20)	4.21 (4.49)

Table 1B – MS cohort used for the direct comparison of QyScore® WMH_U-NET algorithm with six state-of-the-art automated WMH segmentation algorithms, with a training and testing split for those where optimization training was possible.

MS database for algorithm comparison	N of subjects	Clinical status	Age range (mean +- std) (range)	Sex (M - F)
Global	30	24 RRMS, 2 SPMS, 1 PPMS, 2 CIS, 1 unspecified	39.27 +- 10.12 (25 – 64)	7 – 23
Training	10	9 RRMS, 1 SPMS	42.3 +- 11.13 (30 – 64)	1 - 9
Testing	20	15 RRMS, 1 SPMS, 1 PPMS, 2 CIS, 1 unspecified	37.75 +- 9.51 (25 – 60)	6 – 14

HC = Healthy Controls; **AD** = Alzheimer's Disease; **FTD** = Frontotemporal Dementia; **MS** = Multiple Sclerosis; **N/A** – Clinical status not available;
RRMS = relapsing-remitting multiple sclerosis; **SPMS** = secondary progressive MS; **PPMS** = primary progressive MS; **CIS** = clinically isolated syndrome

RESULTS

QyScore® WMH_U-Net demonstrated good volume and spatial overlap (average DSC=0.66±0.2), especially with larger WMH load (15-30ml: DSC=0.75±0.07) across the full validation cohort. Compared to available state-of-the-art algorithms, QyScore® WMH_U-Net outperformed both unsupervised and supervised methods (default settings), producing segmentations most closely matching the consensus manual expert gold-standard (Figure 1B, Table 2A).

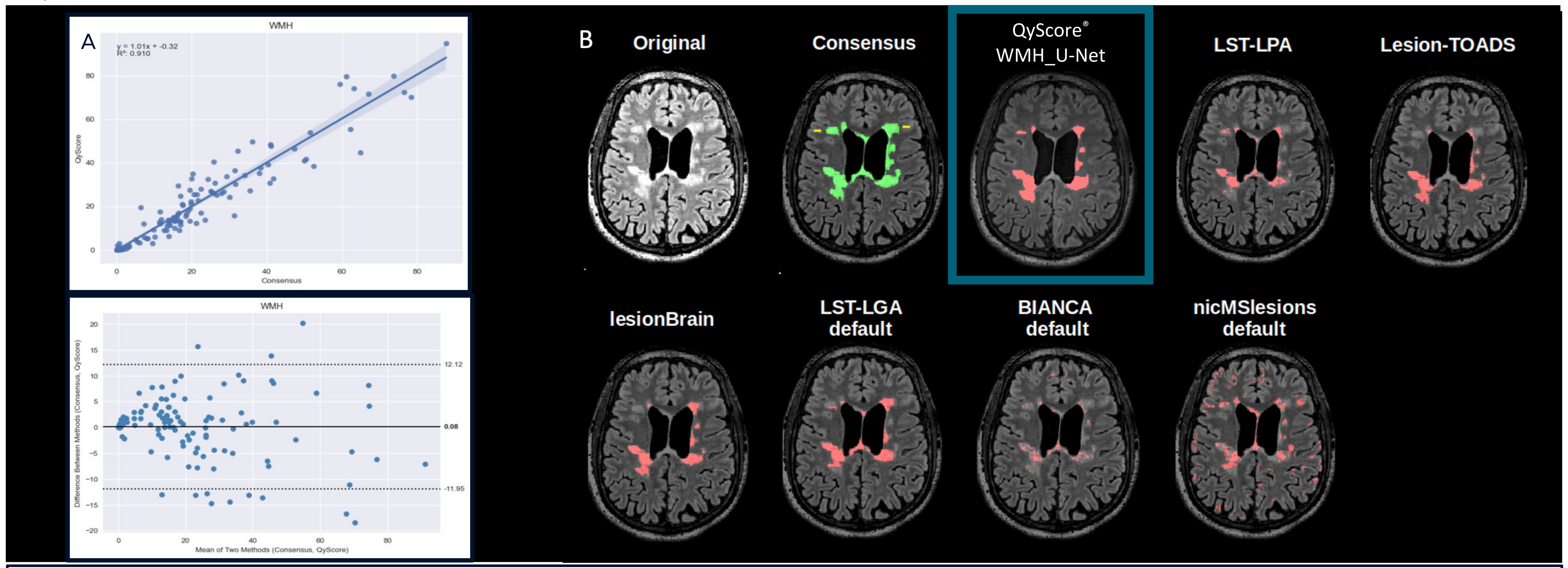

Figure 1A – Linear Regression and Bland-Altman plots demonstrating strong concordance between QyScore® WMH_U-NET and the export manual gold-standard consensus.
Figure 1B – Representative slice demonstrating automated segmentation performance against the expert consensus for QyScore® WMH_U-NET and the six alternative state-of-the-art algorithms.

QyScore® WMH_U-Net demonstrated the highest volume agreement (ICC=0.97) and DSC (0.57±0.24) and lowest AVE (5.33±4.04mL) compared with all 6 segmentation methods (Table 2A and 2B). Optimizing and/or retraining LST-LGA, BIANCA and nicMSlesions on a subset of the MS cohort (Table 2B), improved their performances; however, QyScore® WMH_U-Net remained comparable with the optimized nicMSlesions, and performed better than the optimized LST-LGA and BIANCA. Friedman test (ANOVA) revealed significantly better spatial (DSC: p=5.90x10^{-21}) and volumetric agreement (AVE: p=3.61x10^{-6}) between QyScore® WMH_U-Net and the other methods (default settings: Table 2A). Wilcoxon signed-rank post-hoc analysis (Bonferroni corrected: p<0.0071 for all default and optimized method comparisons) demonstrated QyScore® WMH_U-Net significantly outperformed all methods across spatial and volumetric agreement with an expert manual consensus gold-standard comparator.

Table 2A – Performance metrics for QyScore® WMH_U-NET and six state-of-the-art automated WMH segmentations methods applied to the full MS set (n=30).

Segmentation Method	Lesion volume	AVE	Dice	F1-score	ICC
Expert Manual Consensus	17.39 ± 16.13 (0.34 – 52.45)	N/A	N/A	N/A	N/A
QyScore® WMH_U-Net	12.05 ± 12.97 (0.15 – 41.78)	5.33 ± 4.04 (0.06 – 13.79)	0.57 ± 0.24 (0.08 – 0.86)	0.43 ± 0.15 (0 – 0.63)	0.95
LST-LGA default	8.77 ± 10.06 (0.05 – 36.16)	8.62 ± 7.75 (0.28 – 32.19); *p=3.79E-06**	0.45 ± 0.24 (0.03 – 0.81); *p=2.05E-07**	0.21 ± 0.17 (0.00 – 0.54)	0.83
LST-LPA	5.37 ± 6.37 (0.08 – 23.26)	12.02 ± 10.76 (0.26 – 36.68); *p=5.26E-06**	0.34 ± 0.19 (0.04 – 0.67); *p=9.31E-09**	0.16 ± 0.13 (0.00 – 0.45)	0.61
lesionBrain	7.85 ± 9.52 (0.01 – 32.72)	9.54 ± 7.83 (0.33 – 25.33); *p=3.73E-09**	0.41 ± 0.24 (0.00 – 0.76); *p=3.73E-09**	0.19 ± 0.13 (0.00 – 0.59)	0.83
Lesion-TOADS	15.27 ± 8.31 (3.46 – 36.85)	9.20 ± 6.82 (0.06 – 25.31); *p=4.66E-03**	0.41 ± 0.25 (0.02 – 0.73); *p=1.86E-09**	0.23 ± 0.09 (0.09 – 0.40)	0.61
BIANCA default	2.16 ± 1.62 (0.31 – 36.85)	14.56 ± 14.14 (0.05 – 45.17); *p=3.27E-02**	0.24 ± 0.09 (0.07 – 0.42); *p=3.54E-08**	0.11 ± 0.09 (0.00 – 0.36)	0.25
nicMSlesions default	36.41 ± 24.87 (13.89 – 115.71)	19.89 ± 17.83 (0.32 – 74.97): *p=1.60E-05**	0.18 ± 0.13 (0.00 – 0.41); *p=1.86E-09**	0.07 ± 0.05 (0.00 – 0.18)	0.60

Table 2B – Performance metrics for QyScore® WMH_U-NET and six state-of-the-art automated methods applied to the testing set (n=20) following optimization training. Six default and three possible optimized results presented.

Segmentation Method	Lesion volume	AVE	Dice	F1-score	ICC
Expert Manual Consensus	17.53 ± 17.09 (0.34 – 52.45)	N/A	N/A	N/A	N/A
QyScore® WMH_U-Net	12.67 ± 13.73 (0.02 – 41.79)	3.57 ± 3.52 (0.33 – 13.9)	0.56 ± 0.26 (0.09 – 0.86)	0.42 ± 0.17 (0 – 0.63)	0.97
LST-LGA default	8.60 ± 10.75 (0.02 – 33.82)	8.93 ± 7.39 (0.30 – 20.19) *p*<0.0071	0.41 ± 0.28 (0.00 – 0.78); *p*<0.0071	0.16 ± 0.15 (0.00 – 0.52)	0.86
LST-LGA optimized	15.10 ± 15.23 (1.04 – 46.81)	4.28 ± 3.66 (0.08 – 11.54); *p*=0.388**	0.51 ± 0.26 (0.06 – 0.85); *p*<0.0071**	0.20 ± 0.14 (0.03 – 0.51)	0.95
BIANCA default	2.68 ± 2.22 (0.39 – 7.28)	14.88 ± 15.11 (0.05 – 45.17) *p*<0.0071	0.22 ± 0.08 (0.07 – 0.36); *p*<0.0071	0.09 ± 0.08 (0.00 – 0.32)	0.23
BIANCA optimized	10.90 ± 7.92 (2.88 – 31.39)	8.54 ± 8.55 (0.78 – 31.73); *p*<0.0071**	0.39 ± 0.18 (0.07 – 0.66); *p*<0.0071**	0.23 ± 0.11 (0.07 – 0.42)	0.71
nicMSlesions default	39.79 ± 29.73 (13.89 – 115.71)	22.60 ± 20.93 (0.58 – 74.97) *p*<0.0071	0.17 ± 0.14 (0.00 – 0.41); *p*<0.0071	0.06 ± 0.05 (0.00 – 0.15)	0.61
nicMSlesions optimized	14.33 ± 13.09 (0.00 – 36.90)	4.65 ± 6.78 (0.05 – 27.97); *p*=0.202**	0.63 ± 0.23 (0.00 – 0.85), *p*=0.083**	0.56 ± 0.21 (0.00 – 0.86)	0.88

* Wilcoxon signed-rank tests comparing the DSC and AVE for each of the **six** state-of-the-art methods (default settings) with QyScore® WMH_U-Net segmentations in the full cohort (n=30).
** Wilcoxon signed-rank tests comparing the DSC and AVE for each of the **three** state-of-the-art methods that allowed retraining and optimization with QyScore® WMH_U-Net segmentations.

CONCLUSIONS

QyScore® WMH_U-Net fully-automated WMH segmentations significantly outperformed 6 widely used state-of-the-art automated WMH segmentation tools across multiple spatial and volume performance metrics. It produced fast, robust and accurate WMH segmentations across a varied cohort from multiple scanners and patient groups, supporting its widespread application for clinical routine practice.