

VARIATION IN PERFORMANCE OF DIFFERENT VERSIONS OF A COMMERCIAL MACHINE LEARNING-BASED TOOL FOR THE LONGITUDINAL ANALYSIS OF MULTIPLE SCLEROSIS PLAQUES ON BRAIN MRI

J L Chen, S Kitaoka, S Nesar, M A Esmeraldo, F Zaccagna, T Das
Department of Radiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

BACKGROUND

In recent years, there has been rapid development of artificial intelligence (AI) tools within radiology, including for the detection and characterisation of demyelinating lesions in multiple sclerosis (MS).¹

Although many tools have been clinically evaluated initially, the impact of periodic version updates is unclear without careful re-evaluation. We aimed to evaluate the performance of a MDR Class IIa tool (Pixyl.Neuro.MS) across different versions.^{2,3}

METHODS

We performed a retrospective 3-month review of MRI brain studies for the monitoring of patients with MS.

The performance of Pixyl.Neuro.MS [versions 2.1.0 and 3.0.0] in processing, detecting, and characterising new lesions on 3D T2-FLAIR MRI sequences was assessed on the same set of studies, with presence of any new lesion taken to indicate active disease.⁴

Where the Pixyl.Neuro.MS and consultant neuroradiologist reports were concordant, these were accepted as the ground truth. Where they were discordant, ground truth was established by researcher review.

RESULTS

43 cases (MRI study pairs) were evaluated (patient age range 17–72, mean age 42; male : female ratio = 1 : 1.39; mean scan interval 16.5 months).

The performance of Pixyl.Neuro.MS markedly varied between the versions [2.1.0 → 3.0.0]. Technical processing performance was reduced [43 → 28 cases successfully processed].

Analysis of the 28 cases that were processed by both versions revealed that diagnostic performance also varied. Paired comparison revealed similar overall accuracy [89% → 86%] but comprising markedly decreased sensitivity [100% → 70%] and increased specificity [82% → 94%].

CONCLUSION

Pixyl.Neuro.MS is a promising tool for the longitudinal analysis of demyelinating lesions in multiple sclerosis, particularly for the identification of new lesions (signifying active disease) or for confirmation of no new lesions (stable disease).

However, re-evaluation of Pixyl.Neuro.MS after version update from 2.1.0 to 3.0.0 revealed reduced technical processing performance and altered diagnostic performance. Despite similar overall accuracy, the decreased sensitivity and increased specificity of the newer version may ultimately result in reduced clinical utility, if mainly used in radiological practice as an aid for detecting subtle lesions

These findings may generalise to other AI tools, and we advocate careful post-market surveillance. Utilising a standard set of reference cases for repeated evaluation across versions may be able to reveal performance changes with smaller sample size than would otherwise be necessary.

REFERENCES

1. Mendelsohn, Z. *et al.* Commercial volumetric MRI reporting tools in multiple sclerosis: a systematic review of the evidence. 2023. *Neuroradiology*, 65(1), 5-24.

2. Roca, P. *et al.* Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. 2020. *Diagnostic and Interventional Imaging*, 101(12), 795-802.

3. Pixyl.Neuro.MS. <https://pixyl.ai/multiple-sclerosis/>

4. Giovannoni, G, et al. "Is it time to target no evident disease activity (NEDA) in multiple sclerosis?." *Multiple Sclerosis and Related Disorders* 4.4 (2015): 329-333.

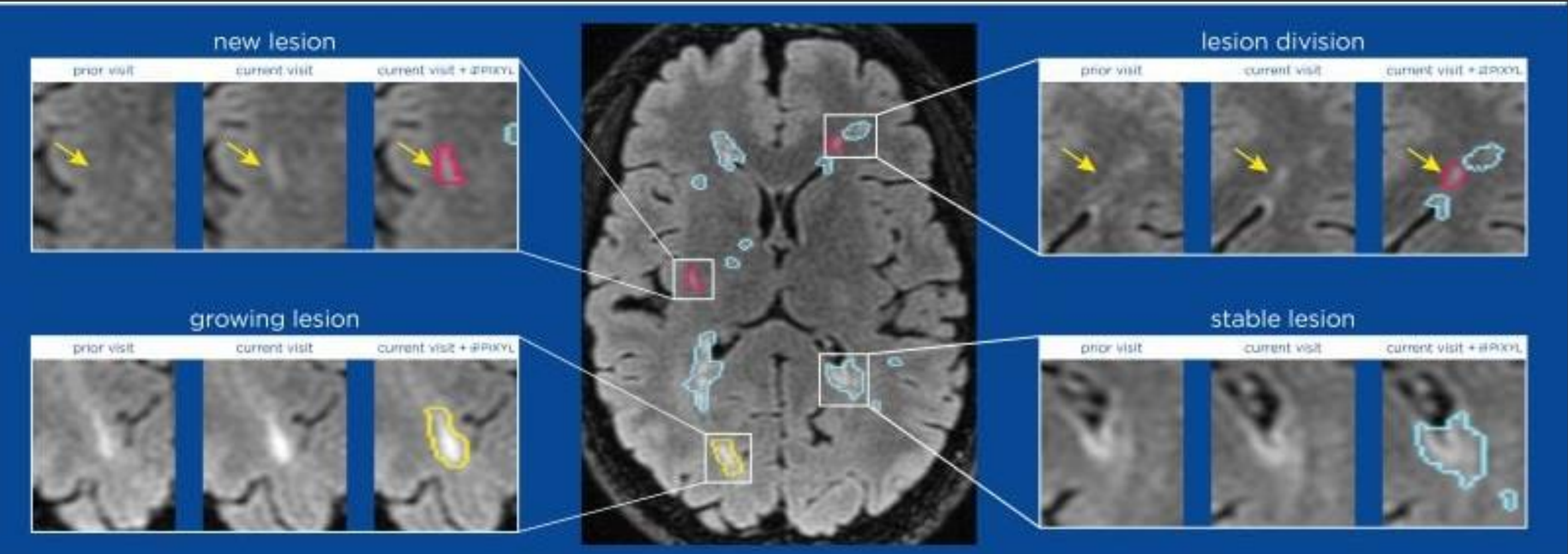


Figure 1. Pixyl.Neuro.MS detection and characterisation of multiple sclerosis lesions as new (red), enlarging (yellow), or stable (blue).

	Pixyl.Neuro.MS 2.1.0	Pixyl.Neuro.MS 3.0.0	Neuroradiologist
Actual +	11	10	11
Actual -	17	18	17
Test +	14	8	12
Test -	14	20	16
Sensitivity	1	0.7	0.909
Specificity	0.823529412	0.944444444	0.882
PPV	0.785714286	0.875	0.833
NPV	1	0.85	0.9375
Accuracy	0.892857143	0.857142857	0.892857143

Figure 2. Performance of Pixyl.Neuro.MS versions 2.1.0 and 3.0.0 and neuroradiologists on the set of cases that were successfully processed by both versions.

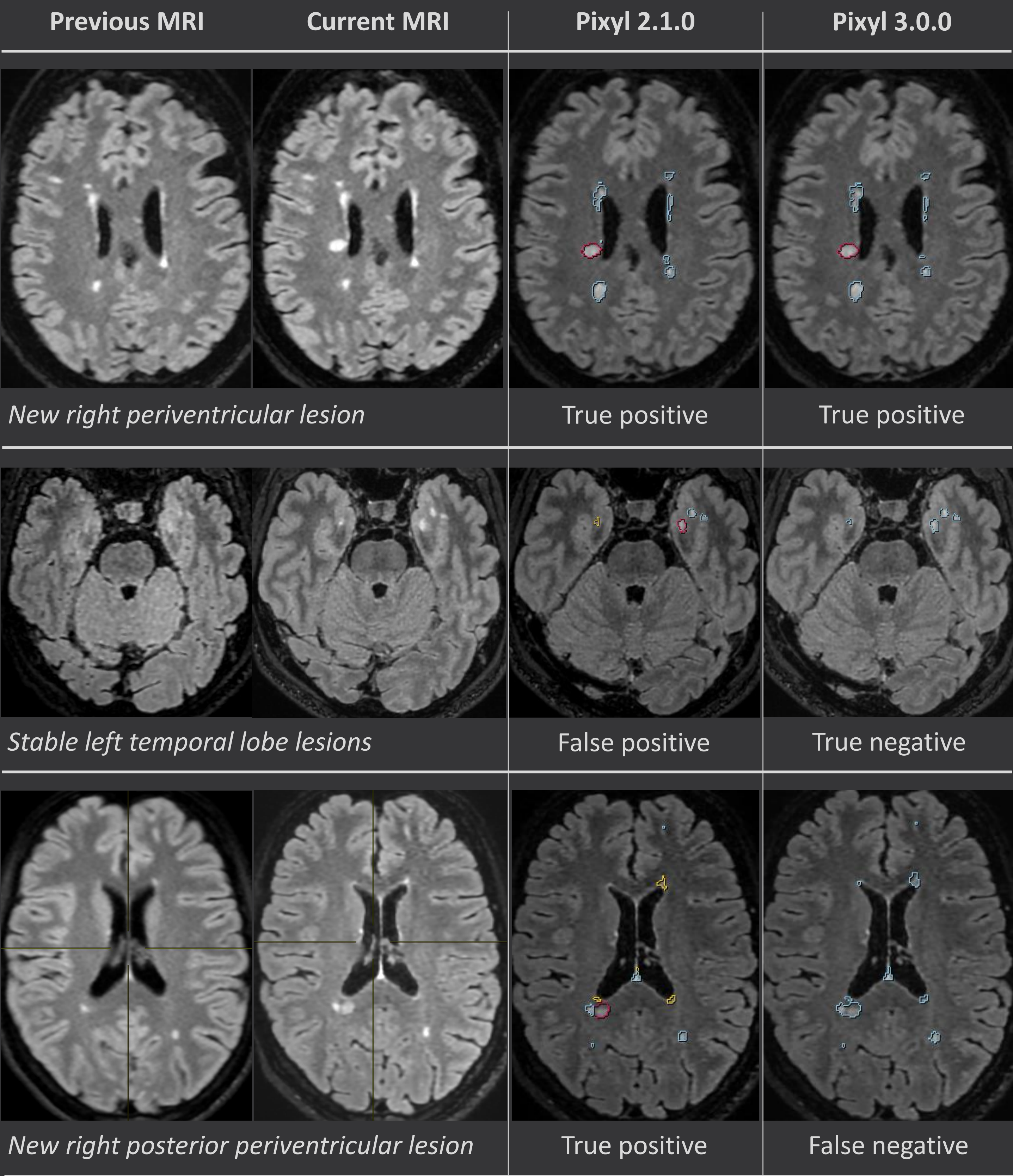


Figure 3. Examples of concordant and discordant results between Pixyl.Neuro.MS versions 2.1.0 and 3.0.0 on the same cases, with an overall increase in specificity but decrease in sensitivity from version 2.1.0 to 3.0.0.