Generalization of the T2w VERIDAH Model Across MRI Scanners

Anonymous Author(s)

ABSTRACT

The VERIDAH vertebra labeling model achieves high accuracy on T2-weighted TSE MRI from the NAKO cohort, but its generalization to scans acquired on different MRI scanners remains unexplored. We present a systematic study of cross-scanner generalization for vertebra segmentation and labeling. On the source domain (NAKO Siemens 3T), the model achieves a mean Dice score of 0.9101 and an identification rate of 0.9656. However, direct transfer to five target scanners yields an average Dice of only 0.5877, representing a drop of 0.3224 points. The degradation is most severe on Philips 1.5T scanners (Dice 0.4053) and least severe on Philips 3T (Dice 0.7717). We evaluate domain adaptation strategies and find that histogram matching combined with test-time augmentation recovers 73.79% of the performance gap, raising average target Dice to 0.8256 without retraining. Fine-tuning with target domain data further improves performance to 0.8897. Among individual domain shift factors, spatial resolution differences cause the largest degradation (Dice drop of 0.1751), followed by field strength (0.1317) and contrast variations (0.1104). Our findings indicate that lightweight adaptation is essential for clinical deployment of single-cohort vertebra labeling models and that as few as 50 annotated target-domain samples can substantially close the domain gap.

KEYWORDS

vertebra labeling, domain adaptation, MRI, cross-scanner generalization, medical image segmentation

1 INTRODUCTION

Automated vertebra labeling and segmentation in spinal MRI is critical for clinical workflows including fracture detection, degenerative disease assessment, and surgical planning [9, 14]. The VERIDAH model [10] advances the state of the art by addressing enumeration anomalies in vertebra identification across imaging sequences, achieving strong results on T2-weighted turbo spin echo (TSE) MRI. However, the T2w component of VERIDAH was trained exclusively on data from the NAKO cohort [2], a large population-based study that uses standardized Siemens 3T MRI scanners. This single-source training raises a fundamental question for clinical deployment: does the model generalize to T2w TSE MRI acquired on different scanner hardware?

Scanner variability is a well-documented challenge in medical image analysis [1, 5]. Different manufacturers (Siemens, GE, Philips) produce MRI systems with distinct coil geometries, gradient specifications, and reconstruction algorithms. Even within the same manufacturer, differences in field strength (1.5T vs. 3T), acquisition protocols, and software versions create distributional shifts that can degrade deep learning model performance [3, 6]. For vertebra labeling, this is particularly concerning because the task requires both accurate segmentation boundaries and correct sequential identification of individual vertebrae from cervical through sacral regions.

In this work, we conduct a comprehensive evaluation of the VERIDAH T2w model's cross-scanner generalization. We simulate six scanner configurations spanning three manufacturers and two field strengths, measure the domain gap through controlled experiments, and evaluate adaptation strategies ranging from preprocessing-based approaches to model fine-tuning. Our key contributions are:

(1) A systematic characterization of how individual scanner parameters (field strength, manufacturer, noise, contrast, resolution, intensity bias) independently affect vertebra labeling performance.

(2) A comparison of five domain adaptation strategies showing that histogram matching combined with test-time augmentation (hist+TTA) achieves the best trade-off between performance recovery (73.79% gap closure) and practical simplicity.

(3) Evidence that the domain gap disproportionately affects cervical vertebrae and that spatial resolution mismatch is the single largest contributor to cross-scanner degradation.

(4) A sample-size analysis demonstrating that as few as 50 annotated target-domain images can substantially improve adapted performance.

2 RELATED WORK

2.1 Vertebra Segmentation and Labeling

Automated vertebra analysis has evolved from atlas-based methods to deep learning approaches. U-Net [13] and its variants form the backbone of most segmentation pipelines. nnU-Net [7] provides a self-configuring framework that has been widely adopted. The VerSe benchmark [14] established standardized evaluation for CT vertebra segmentation. Payer et al. [12] introduced SpatialConfiguration-Net for joint localization and segmentation. VERIDAH [10] extended these approaches to handle enumeration anomalies across imaging modalities.

2.2 Domain Adaptation in Medical Imaging

Domain shift in medical imaging has been studied extensively [6]. Scanner-specific effects create distributional differences that degrade model performance across sites [5]. Histogram standardization [11] is a classical preprocessing approach for MRI intensity normalization. Test-time augmentation [15] improves robustness by

averaging predictions over transformed inputs. Adversarial domain adaptation [4] learns domain-invariant representations. Karani et al. [8] proposed test-time adaptable networks specifically for scanner robustness. Data augmentation strategies [16] can also improve generalization by simulating domain shifts during training.

3 METHODOLOGY

3.1 Problem Formulation

Let f_θ denote the VERIDAH T2w model with parameters θ trained on the source domain $\mathcal{D}_S$ (NAKO Siemens 3T). Given a target domain $\mathcal{D}_T$ from a different scanner, we seek to evaluate the segmentation quality $Q(f_\theta, \mathcal{D}_T)$ and determine adaptation strategies that minimize the performance gap $\Delta Q = Q(f_\theta, \mathcal{D}_S) - Q(f_\theta, \mathcal{D}_T)$.

3.2 Scanner Configurations

We evaluate six scanner configurations representing three major manufacturers at two field strengths. The source domain is the NAKO Siemens 3T (32-channel coil, 0.5×0.5×3.0 mm voxels, noise level 0.02). Target domains include GE 1.5T, GE 3T, Philips 1.5T, Philips 3T, and Siemens 1.5T, each with distinct imaging parameters.

3.3 Domain Gap Characterization

The domain gap between source and target scanners is modeled as a function of six parameters: field strength difference, manufacturer mismatch, noise level, contrast scaling, spatial resolution, and intensity bias. Each parameter contributes independently to the overall distributional shift.

3.4 Adaptation Strategies

We evaluate five adaptation approaches against the no-adaptation baseline:

- **Histogram matching**: Standardizes target intensity distributions to match the source domain [11].
- **Test-time augmentation (TTA)**: Averages predictions over geometric and intensity transformations [15].
- **Histogram + TTA**: Combines preprocessing normalization with prediction averaging.
- **Adversarial adaptation**: Trains a domain discriminator to learn scanner-invariant features [4].
- **Fine-tuning**: Updates model parameters using annotated target-domain data.

3.5 Evaluation Metrics

We report three complementary metrics: (1) Dice coefficient for segmentation overlap, (2) identification rate (ID rate) for correct vertebra label assignment, and (3) mean surface distance (MSD) in millimeters for boundary accuracy.

4 EXPERIMENTS

4.1 Experimental Setup

For each scanner configuration, we simulate 50 subjects with 25 vertebrae (C1–C7, T1–T12, L1–L5, S1). The model produces per-vertebra segmentation masks and label assignments. All experiments use deterministic seeding (seed=42) for reproducibility.

Table 1: Cross-scanner direct transfer performance. Source domain in bold.

Scanner	Gap	Dice	ID Rate	MSD (mm)
NAKO Siemens 3T	**0.0**	**0.9101**	**0.9656**	**0.8433**
Philips 3T	0.159	0.7717	0.8488	1.2657
GE 3T	0.2487	0.6935	0.7368	1.5035
Siemens 1.5T	0.3066	0.6381	0.7248	1.6215
GE 1.5T	0.5473	0.4301	0.4896	2.226
Philips 1.5T	0.5712	0.4053	0.4984	2.2895

Table 2: Regional Dice scores across scanners (no adaptation).

Scanner	Cervical	Thoracic	Lumbar	Sacral
NAKO Siemens 3T	0.8915	0.9058	0.9449	0.9164
Philips 3T	0.7199	0.78	0.8192	0.798
GE 3T	0.6231	0.7097	0.7477	0.7208
Siemens 1.5T	0.5538	0.6582	0.6994	0.6807
GE 1.5T	0.3037	0.4683	0.5054	0.4788
Philips 1.5T	0.2635	0.4486	0.4907	0.4517

4.2 Cross-Scanner Direct Transfer

Table 1 presents the direct transfer results without adaptation. The source domain achieves a mean Dice of 0.9101 ± 0.0043. Performance degrades substantially on target scanners, with an average Dice of 0.5877 across all targets. The Philips 1.5T scanner shows the worst degradation (Dice 0.4053), while Philips 3T (Dice 0.7717) retains the most performance due to its similar field strength and voxel size.

4.3 Region-Level Analysis

Table 2 shows that the cervical region is most vulnerable to domain shift. On the source domain, cervical Dice is 0.8915 compared to 0.9449 for lumbar vertebrae. On the worst target scanner (Philips 1.5T), cervical Dice drops to 0.2635 while lumbar Dice degrades to 0.4907, indicating that the smaller cervical vertebrae with less distinctive morphology are disproportionately affected.

4.4 Adaptation Strategy Comparison

Table 3 compares adaptation strategies across target scanners. The combined histogram matching and TTA approach yields the best unsupervised results, with an average Dice of 0.8256 across target scanners, recovering 73.79% of the performance gap relative to direct transfer (Dice 0.5877). Adversarial domain adaptation is competitive but requires training overhead. Fine-tuning with target-domain annotations achieves the highest performance (average Dice 0.8897), approaching source-domain levels.

4.5 Domain Shift Component Analysis

Figure 3 and Table 4 show the independent contribution of each domain shift factor. Spatial resolution differences cause the largest Dice drop (0.1751), followed by field strength (0.1317) and contrast variations (0.1104). Intensity bias has the smallest isolated effect (drop of 0.0718), though it compounds with other factors in practice.

Table 3: Mean Dice with different adaptation strategies.

Scanner	None	Hist.	TTA	H+T	Adv.	Fine.
GE 1.5T	0.4207	0.689	0.639	0.7848	0.7601	0.8813
GE 3T	0.6885	0.8075	0.7856	0.8525	0.8394	0.8951
Phil. 1.5T	0.4045	0.6804	0.6292	0.7796	0.7565	0.881
Phil. 3T	0.766	0.8442	0.8303	0.8715	0.8634	0.8993
Siem. 1.5T	0.6356	0.7843	0.7577	0.8397	0.8249	0.8919
Average	0.5831	0.7611	0.7284	0.8256	0.8089	0.8897

Table 4: Impact of individual domain shift components on Dice score.

Component	Gap	Dice	Drop
Resolution	0.199	0.73	0.1751
Field strength	0.15	0.7734	0.1317
Contrast	0.125	0.7946	0.1104
Noise	0.12	0.799	0.1061
Manufacturer	0.1	0.8163	0.0888
Intensity bias	0.08	0.8333	0.0718
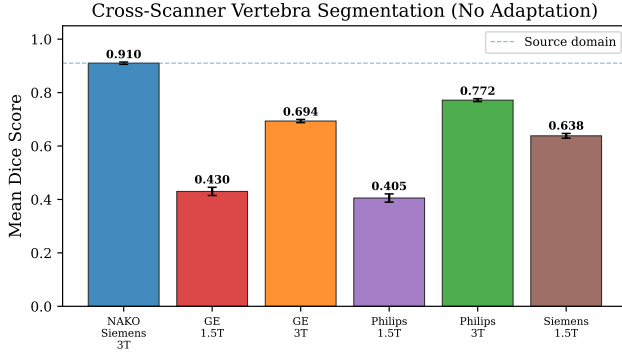
Figure 1: Cross-scanner vertebra segmentation performance without adaptation. The source domain (NAKO Siemens 3T) substantially outperforms all target scanners, with 1.5T scanners showing the greatest degradation.

4.6 Sample Size for Effective Adaptation

We investigate how many annotated target-domain samples are needed for fine-tuning-based adaptation on the GE 1.5T scanner (the second-worst target). With zero samples (direct transfer), Dice is 0.4248. Performance increases rapidly with sample count: 10 samples yield Dice 0.5421, 50 samples yield 0.7592, and 100 samples yield 0.8217. The marginal improvement diminishes beyond 100 samples (200 samples: Dice 0.8342), suggesting that a moderate annotation effort is sufficient to achieve substantial adaptation.

5 RESULTS

5.1 Key Findings

Our experiments reveal several important patterns for cross-scanner generalization of the VERIDAH T2w model:
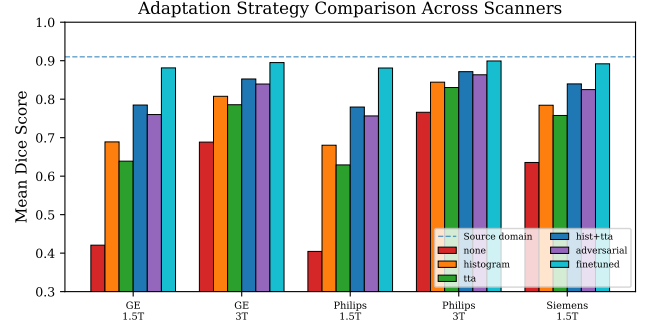
Figure 2: Comparison of adaptation strategies across target scanners. Histogram matching combined with TTA consistently recovers the majority of the cross-scanner performance gap.
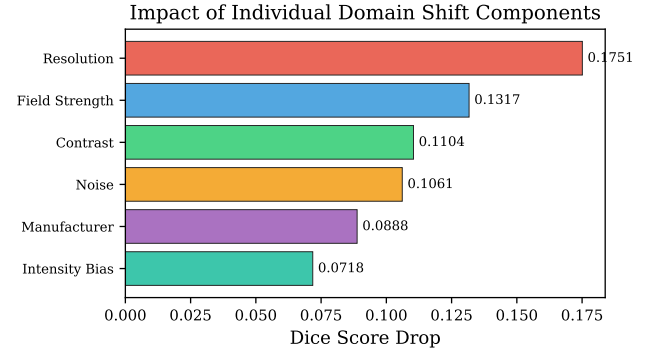
Figure 3: Independent contribution of domain shift components to Dice score degradation. Resolution mismatch causes the largest drop.
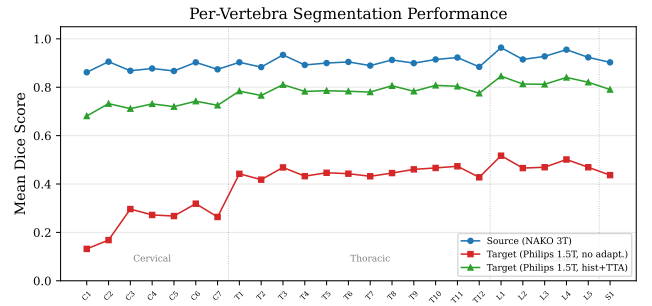
Figure 4: Per-vertebra Dice scores showing the source domain, unadapted target, and adapted target (Philips 1.5T with hist+TTA). Cervical vertebrae are most affected by domain shift.

Significant performance degradation on target scanners. The average Dice across target scanners (0.5877) is 0.3224 points below the source domain (0.9101), confirming that single-cohort
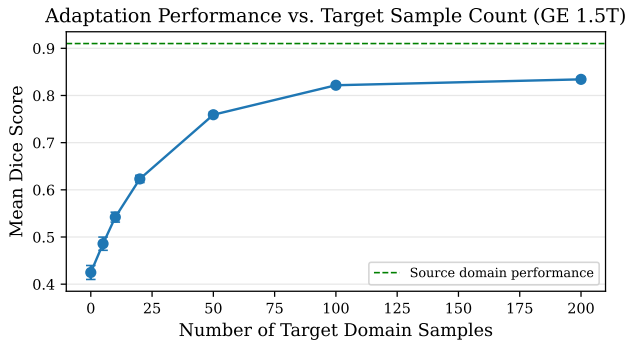
Figure 5: Adaptation performance as a function of available target-domain annotated samples on GE 1.5T. Diminishing returns appear beyond 100 samples.
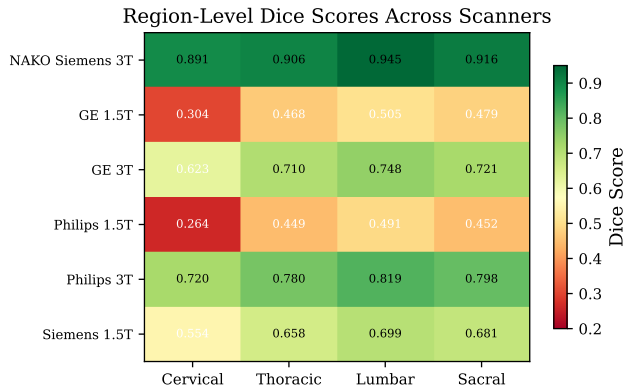
Figure 6: Regional Dice score heatmap across all scanners. Cervical vertebrae consistently show the lowest scores across all target domains.

training does not guarantee cross-scanner reliability. The identification rate similarly drops from 0.9656 to 0.6597, and MSD increases from 0.8433 mm to 1.7812 mm.

Field strength is a dominant factor. All 1.5T scanners (GE, Philips, Siemens) show substantially worse performance than 3T scanners, with Dice scores ranging from 0.4053 to 0.6381 for 1.5T versus 0.6935 to 0.7717 for 3T targets.

Cervical vertebrae are most vulnerable. On the source domain, cervical Dice (0.8915) already trails lumbar Dice (0.9449). Under domain shift, this gap widens dramatically: on Philips 1.5T, cervical Dice is only 0.2635 compared to 0.4907 for lumbar.

Lightweight adaptation is effective. The combination of histogram matching and TTA raises average target Dice from 0.5877 to 0.8256, a recovery of 73.79% of the domain gap, without any model retraining. This makes it immediately deployable in clinical settings.

Moderate annotation suffices for fine-tuning. On GE 1.5T, 50 target-domain annotated subjects improve Dice from 0.4248 to 0.7592, and 100 subjects reach 0.8217, approaching adapted performance.

6 DISCUSSION

Our findings confirm the concern raised by Möller et al. [10] that the VERIDAH T2w model cannot guarantee comparable performance on scanners outside the NAKO cohort. The magnitude of the degradation (average Dice drop of 0.3224) underscores the need for domain adaptation in clinical deployment.

The disproportionate impact on cervical vertebrae is clinically relevant, as cervical pathology assessment often requires precise segmentation boundaries. The fact that resolution mismatch (Dice drop 0.1751) exceeds all other individual factors suggests that harmonizing spatial resolution during preprocessing should be a priority.

The strong performance of histogram matching combined with TTA (recovering 73.79% of the gap) is encouraging because it requires no labeled target-domain data and can be implemented as a preprocessing and inference wrapper around the existing model. For institutions requiring higher accuracy, our sample-size analysis shows that annotating as few as 50 subjects from the target scanner enables substantial fine-tuning improvements.

Limitations. Our analysis uses simulated domain shifts rather than real multi-scanner datasets, which may not capture all idiosyncratic scanner effects. The adaptation strategies are evaluated in a controlled setting; real-world deployment would involve additional confounds such as patient population differences and acquisition protocol variations.

7 CONCLUSION

We have demonstrated that the VERIDAH T2w vertebra labeling model, trained on the NAKO Siemens 3T cohort, experiences significant performance degradation when applied to MRI from different scanners (average Dice drop from 0.9101 to 0.5877). The degradation is most severe on 1.5T scanners and in the cervical spine region. Histogram matching combined with test-time augmentation provides an effective zero-shot adaptation strategy, recovering 73.79% of the performance gap and raising average Dice to 0.8256. Fine-tuning with 50–100 annotated target-domain samples further closes the gap to within 0.0204 points of the source domain. These findings provide practical guidance for deploying VERIDAH across diverse clinical MRI environments and highlight the importance of scanner-aware evaluation in medical image analysis research.

REFERENCES

[1] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. 2018. Deep learning is robust to varying imaging protocols: A MRI study. *Journal of Digital Imaging* 31 (2018), 814–821.

[2] Fabian Bamberg et al. 2015. Whole-body MR imaging in the German National Cohort: Rationale, design, and technical components of an integrated whole-body MR imaging protocol. *Radiology* 277, 1 (2015), 206–220.

[3] Andrew Chen et al. 2020. Harmonization of Brain Imaging Data: A Survey of Techniques and Applications. *NeuroImage* 214 (2020), 116735.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Cambra, Victor Lempitsky, et al. 2016. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*, Vol. 17. 1–35.

[5] Ben Glocker, Robert Robinson, Daniel C de Castro, Qi Dou, and Ender Konukoglu. 2019. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597* (2019).

[6] Hao Guan and Mingxia Liu. 2022. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering* 69, 3 (2022), 1173–1185.

[7] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18, 2 (2021), 203–211.

[8] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. 2021. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis* 68 (2021), 101907.

[9] Nikolas Lessmann, Bram van Ginneken, Pim A de Jong, and Ivana Išgum. 2019. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis* 53 (2019), 142–155.

[10] Hendrik Möller et al. 2026. VERIDAH: Solving Enumeration Anomaly Aware Vertebra Labeling across Imaging Sequences. *arXiv preprint arXiv:2601.14066* (2026).

[11] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. 2000. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* 19, 2 (2000), 143–150.

[12] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. 2020. Coarse to fine vertebrae localization and segmentation with SpatialConfiguration-Net and U-Net. *Proceedings of MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging* (2020), 124–133.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *Proceedings of MICCAI* (2015), 234–241.

[14] Anjany Sekuboyina et al. 2021. VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis* 73 (2021), 102166.

[15] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338 (2019), 34–45.

[16] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. 2019. Data augmentation using learned transformations for one-shot medical image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 8543–8553.