Studies in Knowledge and Innovation
Vol. 2 No. 10A (2026)

From Vascular Masks to Structural Conclusions: Evaluating MorVess across Acquisition Contexts

Julian Peterson[1] · Nathan Marshall[2] · Matthew Fisher[3]

[1] Mexico · [2] Bahrain · [3] Cuba

Published: October 26, 2026 | Article 930

Abstract

Acquisition differences can change the measurement problem faced by a segmentation model. Transport across image sources should therefore be read as a comparison of maintained assumptions, not only a new score.

From Vascular Masks to Structural Conclusions: Evaluating MorVess across Acquisition Contexts

Abstract

Acquisition differences can change the measurement problem faced by a segmentation model. Transport across image sources should therefore be read as a comparison of maintained assumptions, not only a new score.

Introduction

A structural conclusion drawn from a segmentation mask depends partly on the images from which that mask was produced. For the topic of Mao et al. (2026), differences between acquisition contexts may alter visibility, contrast, or the representation of small structures. A review should identify which aspect of the original segmentation claim is being extended when a new image source is considered. A changed aggregate score cannot by itself distinguish model weakness from a changed measurement target. A careful comparison retains both the source context and the anatomical property of interest. Transfer is then discussed through what remained comparable and what changed, rather than through one universal accuracy label.

From a mask to a scientific conclusion

A segmentation's utility extends beyond its accuracy, especially when it serves as a foundation for downstream analyses or scientific conclusions. Even if a model's errors are significant, they must not compromise the validity of the intended inference. Conversely, a high aggregate score might be misleading if it doesn't align with the specific structural elements of interest. Thus, the evaluation unit should encompass both the segmentation and the downstream measurement to ensure that computational improvements directly contribute to meaningful scientific claims.

The key challenge lies in distinguishing between annotation uncertainty and model error, as well as recognizing cases where boundaries cannot be resolved confidently. These ambiguous regions highlight issues within the observation process rather than flaws in the dataset itself. By framing the evaluation in this way, we can leverage computational progress to support precise scientific conclusions: the model enhances a defined measurement under specified conditions, while broader structural and clinical implications remain distinct. This approach allows for nuanced discussions about how different acquisition contexts affect the measurement problem faced by segmentation models.

Biomedical image representations

The biomedical segmentation model, exemplified by U-Net, operates within a structured framework, generating an inferred representation whose meaning is contingent upon the annotation conventions and intended application (Ronneberger et al., 2015). This schematic illustrates that the choice of input data, preprocessing steps, and training conditions collectively form a comprehensive pipeline. When comparing models like nnU-Net, it's crucial to acknowledge the specific pipeline used in each assessment. Ignoring these details can lead to an incomplete understanding of the model's performance across different acquisition contexts, as suggested by the review question about decision-making and independence from selection (Kapoor & Narayanan, 2023). Thus, evaluating a segmentation model requires not only assessing the model architecture but also considering the full pipeline of choices made during image representation and training.

Agreement needs an account of independence

Agreement in a study can arise from two distinct sources: independent observations or shared assumptions that are inherited across different data sets. Both types of agreement provide valuable insights, but they serve different purposes. Independent observations offer robust support by showing that results hold true across varied conditions, whereas shared assumptions indicate consistency among different descriptions of the same phenomenon.

To better understand these distinctions, a future synthesis could categorize agreements into shared observations and shared representations. This categorization would help clarify whether an alternative method or context genuinely challenges a conclusion or merely repeats the same link in the argument. Discordance between these categories becomes informative because it reveals which assumptions are not consistently maintained across different contexts.

The goal is not to dismiss agreement outright but to ensure that conclusions are robust and well-supported by independent evidence. Identifying which assumptions persist across different datasets helps validate broader findings, ensuring that conclusions are not merely repetitions of the same link in an argument.

The complete pipeline is the reproducible object

The comparison depends on ensuring that the complete observation pipeline is reproducible and documented. As Cardoso et al. (2022) emphasize, a network's name does not fully encapsulate the pipeline's full process, which can include preprocessing, implementation choices, data handling, sampling, training, and selection decisions. These steps are crucial as they determine the model's performance and can vary significantly across different acquisition contexts, leading to varying measurement outcomes.

Furthermore, Kapoor and Narayanan (2023) highlight that repeated adjustments of a pipeline after inspecting the same evaluation data can undermine its independence. This practice transforms the data into development evidence rather than independent assessment, potentially diluting the validity of the comparison. A clear distinction between development choices and independent evaluation is necessary to maintain the integrity of the scientific contribution. By documenting each step of the pipeline, researchers ensure that their findings are reproducible and not influenced by iterative adjustments.

Geometry and metric selection

Metrics Reloaded underscores the importance of selecting metrics that align with the specific challenges and goals of a given imaging task. When transitioning from one dataset to another, it is crucial to maintain consistent assumptions about the structure and continuity of the data. For instance, clDice, while effective for assessing tubular structures, emphasizes continuity and may not reflect anatomical correctness, which can be a significant issue when the task requires precise structural integrity.

On the other hand, metrics Reloaded advocates for specifying the intended target and the consequent error to choose appropriate metrics. This approach allows for a more nuanced understanding of the structure and its implications across different imaging contexts. For example, if the goal is to evaluate the continuity of blood vessels, using clDice might be sufficient, but when assessing the structural integrity of vascular walls, a metric that accounts for the specific geometry and topology of these structures becomes essential.

While metrics like clDice can provide valuable insights into the continuity of tubular structures, they may not capture all aspects of structure, especially in contexts where precise anatomical relationships are

critical. Thus, selecting the right metrics requires careful consideration of the intended scientific conclusions and the specific properties of the imaging data being analyzed.

The decision pathway deserves its own evaluation

The decision pathway deserves its own evaluation, as the context of acquisition differences can alter the measurement challenges faced by segmentation models. A single technical assessment might provide a favorable summary but does not guarantee a complete practical recommendation. As such, evaluating the entire pathway from observation to response is crucial. This evaluation should consider the value and implications of each step in the process, including cases where further review or deferral may be necessary. By doing so, we can distinguish between useful signals and effective decision processes, while keeping the original scientific question in focus. Evaluating the whole pathway ensures that we don’t mistake a technically favorable summary for a complete practical recommendation, thus maintaining the integrity of the scientific inquiry.

Foundation-model adaptation does not erase acquisition differences

In Kirillov et al.'s (2023) foundational work, the segmentation model serves as a starting point for specialized adaptation rather than providing independent evidence that all medical contexts are fully covered. Differences in resolution, contrast, annotation conventions, and sampled anatomy can arise when moving from one acquisition setting to another. These differences imply that models trained in one environment may not maintain the same assumptions or metrics when applied to different imaging sources. Thus, agreement with one annotation process is not necessarily indicative of correspondence with another, especially if the intended use of the model changes. This suggests that independent evaluation of acquisition settings is crucial, as it highlights how the structural quantity of interest can vary significantly between different image sources.

Pipeline and acquisition validity

In the context of evaluating MONAI and UNETR pipelines, it is crucial to recognize that the complete computational pipeline includes more than just the named network. This broader pipeline encompasses a variety of choices such as geometry and metric selection, which are essential for ensuring that the results are valid across different institutions or acquisition environments. The validity of these choices should be preserved and not conflated with software availability alone. For instance, MONAI's framework highlights the importance of maintaining assumptions through the entire pipeline to ensure that measurements remain valid when applied in various settings.

Similarly, UNETR’s use of transformers for three-dimensional medical segmentation underscores the significance of architectural capacity but does not automatically guarantee accuracy. Architectural choices are important, but they must be validated within the specific context of the data and task at hand. The selection of metrics is another critical aspect, as different measures may highlight different types of errors. Therefore, selecting metrics based on the specific task rather than convention alone is essential to accurately assess the performance of segmentation models across varying acquisition environments.

Metric choice follows the consequential error

In selecting metrics, errors such as overlap, boundary distance, and continuity are crucial for understanding different measurement challenges faced by segmentation models. For instance, Maier-Hein et al. (2024) and Taha and Hanbury (2015) highlight that these properties can reveal distinct shortcomings in image analysis. A high score on one of these metrics does not automatically validate the model across all dimensions. It must align with the specific scientific use case. The authors emphasize that the choice of metric is pivotal, especially when different structures contribute unequally to the overall assessment. By ensuring that selected evidence addresses the property preserved by the study, the metric choices can be made interpretable and avoid conflicting results being treated as votes for preferred models.

This ensures that the chosen evidence addresses the specific property the paper aims to preserve, rather than conflating various measurement challenges under a single metric. It is about making the metric choice consequential, aligning it with the intended scientific use of the segmentation model.

Continuity is a preference that needs independent testing

By distinguishing between structural plausibility and anatomical correctness, we can avoid conflating these evaluative objects. The prior serves to guide and constrain the analysis, but it does not inherently validate or settle the uncertainty associated with specific anatomical interpretations.

A vessel representation is an observation model

A vessel representation, as a model of observation, connects the image and the annotated structure through assumptions about boundaries and spatial context. Ronneberger et al. (2015) and Hatamizadeh et al. (2021) provide architectural perspectives that frame these connections. However, while these models are crucial for understanding the anatomy, they do not encapsulate every anatomical feature. If a conclusion hinges on a small branch or a specific local boundary, then relying solely on an average overlap measure may be insufficient. Metrics should align with the specific question at hand, rather than being conventional choices. Thus, the evaluation of these models must make explicit how their consequential errors relate to the questions they address. This ensures that the vessel representation remains interpretable and relevant to its intended purpose.

The independent unit follows the claim

The number of recorded observations does not automatically equate to the number of independent comparisons, which is a critical design consideration. This distinction is particularly important when moving between different acquisition contexts, as it acknowledges that maintaining the same assumptions across datasets can lead to meaningful comparisons rather than simply transferring scores. For instance, if a segmentation model operates under certain geometric and metric assumptions, these should be consistently applied regardless of the source image. A future report could clarify how units are allocated, observed, selected, and analyzed in each context. These units need not be identical but must maintain a clear relationship for interpretation. Detailed observations from one context might excel at characterizing specific features, yet they may lack sufficient variation to fully assess uncertainty in a new context. Thus, while these detailed observations can add value, their volume should not overshadow the independent variation needed to ensure robust conclusions across different acquisition scenarios. Establishing this relationship allows for a more nuanced assessment of the strength and reach of evidence across varied contexts.

What the observation represents

Consider a hypothetical chart with entries representing different acquisition contexts. In one entry, a vessel is segmented into a mask, which captures the vessel's outline and interior structure. At a second level of aggregation, we might consider the same vessel represented by its average properties across multiple acquisitions. Here, the unit represents an average profile or characteristic of the vessel, rather than a single instance. The interpretation shifts from a specific instance (a mask) to a more generalized characterization (average properties). This change reflects the need for comparison not just in terms of new scores but also in maintaining and comparing assumptions across different acquisition contexts. Thus, the unit's representation changes from a singular instance to an average characteristic, requiring a broader account of independence and continuity in the analysis.

Sorting records into comparable groups

When sorting records into comparable groups, it is plausible that different inclusion criteria could alter the appearance of an observed association. For instance, if we imagine a dataset of vascular images, selecting different masks based on varying thresholds might highlight different features, making certain vessels more visible than others. This could skew the results, leading to an association that appears stronger or weaker depending on the specific characteristics used in segmentation. In this hypothetical scenario, the groupings and inclusion criteria would act like filters, altering the composition of the dataset and thus changing the apparent relationship between variables.

The same concept applies to the broader context of image acquisition. If we consider different imaging modalities or conditions, each might emphasize certain aspects of vascular anatomy over others. For example, contrast-enhanced scans might reveal more detailed information about blood flow dynamics than plain scans, leading to a different observed association. Thus, the choice of acquisition context can influence what is seen as significant, akin to how different filters in an image processing pipeline might highlight various features. In this way, sorting records into comparable groups involves not just comparing maintained assumptions but also understanding how these comparisons are framed by the data selection process itself.

Conclusion

One consequence for interpretation is that the shift from a specific instance (like a mask) to a more generalized characteristic (like average properties) requires careful consideration of the underlying assumptions. This leads to a need for a broader account of independence and continuity in the analysis across different acquisition scenarios. An open question is how to ensure that the units used in one context are sufficiently representative and varied enough for meaningful comparison in another context, especially when different acquisition methods introduce varying levels of detail or noise.

References

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., et al. (2024). Metrics reloaded: recommendations for image analysis validation. *Nature Methods, 21(2), 195-212*. https://doi.org/10.1038/s41592-023-02151-z

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., et al. (2022). MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*. https://arxiv.org/abs/2211.02701

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., Maier-Hein, K. H. (2020). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18(2), 203-211*. https://doi.org/10.1038/s41592-020-01008-z

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment Anything. *arXiv preprint arXiv:2304.02643*. https://arxiv.org/abs/2304.02643

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv preprint arXiv:1505.04597*. https://arxiv.org/abs/1505.04597

Taha, A. A., Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging, 15(1), 29*. https://doi.org/10.1186/s12880-015-0068-x

Mao, F., Chen, Y., Wu, B., et al. (2026). MorVess: Morphology-Aware Pulmonary Vessel Segmentation Network. *arXiv preprint arXiv:2606.24214*.

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4(9), 100804*. https://doi.org/10.1016/j.patter.2023.100804

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2021). UNETR: Transformers for 3D Medical Image Segmentation. *arXiv preprint arXiv:2103.10504*. https://arxiv.org/abs/2103.10504