Robust-Kit: Verification tests for AI Robustness

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Artificial intelligence (AI)-driven medical image segmentation has demonstrated remarkable potential and high performance. However, its reliability and trustworthiness in clinical settings are often compromised by domain shifts and out-of-distribution (OOD) data. To address this challenge, we introduce Robust-Kit, a comprehensive framework for evaluating the robustness of AI models in medical image segmentation. Robust-Kit provides a standardized, training-free procedure to assess medical-specific distribution shifts, including acquisition, population, prevalence, and concept shifts. Our framework enables robustness assessment for both models with transparent training data (*Open*) and black-box models with unknown training data (*Closed*). Additionally, we propose two novel metrics: *Robustness Grade (RG)* and *Robustness-Aware Performance (RAP)*, which measure the robustness for *Open* and *Closed* models, respectively. Our key contributions include: (1) A standardized robustness evaluation methodology that eliminates the need for re-training; (2) A comprehensive robustness testing framework tailored for medical image segmentation models; (3) Two novel robustness metrics that enable users to immediately grasp how data shifts impact model performance.
The code is available at https://anonymous.4open.science/r/robust-kit/

Keywords: Robustness · Validation · Trustworthy AI

1 Introduction

Artificial intelligence (AI) has revolutionized medical applications, demonstrating excellence in diagnosis, predictive analytics, and treatment planning [14,6]. However, real-world deployment raises concerns about trust, particularly regarding robustness. Noisy inputs, domain shifts, and out-of-distribution (OOD) data can undermine performance by exploiting model weaknesses, posing significant risks in medical settings [9]. Recent works have addressed robustness to distribution shifts by curating specialized training, validation, and test datasets that allow for robust model evaluation and training [16,13]. While these approaches provide valuable insights, they often necessitate model retraining on specific datasets to compute the drop in performance between training and test datasets. Notably, robustness assessments that are agnostic to the training dataset and independent of the training process remain an open challenge. Robustness concern is particularly critical for Foundation Models (FMs). Due to the vast and

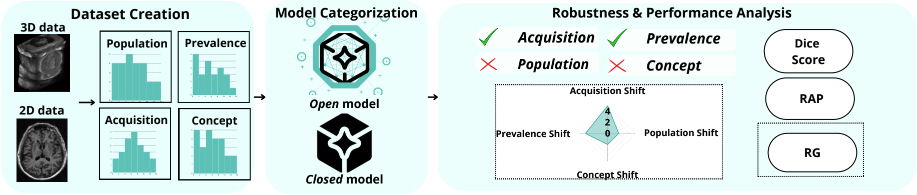

Fig. 1. Overview of Robust-kit. From left to right: Dataset Construction, ensuring the availability of different shifts; Model Categorization, featuring two types of models; and Robustness & Performance Analysis, assessing model performance under varying conditions. The two boxed analyses with dotted lines are only available if the training set is known.

often opaque nature of their training datasets, systematic evaluation is often difficult. Indeed, traditional robustness assessments heavily rely on well-defined statistical properties of the training data [21], and the scale and opacity of FMs, establishing effective robustness measures remains a tough challenge.

Why is Robust-kit needed? To the best of our knowledge, no existing framework offers a standardized pipeline for systematically evaluating robustness against common data shifts in medical image segmentation without the need for re-training the model on a specific set of data. We address this gap by introducing a novel training-free framework for robustness assessment that is broadly applicable across clinical settings. Our contributions can be summarized as follows: (1) **A standardized robustness evaluation methodology** that eliminates the need for model retraining, ensuring efficiency and consistency across assessments. By eliminating the dependency on dataset-specific re-training, Robust-kit establishes a practical approach for evaluating model reliability in diverse segmentation tasks, imaging modalities, anatomical structures, and data shift types. (2) **A comprehensive robustness testing framework** including a systematic suite of robustness tests tailored for medical image segmentation models. The framework consists of three stages illustrated in Figure 1: dataset construction, model categorization, and robustness and performance analysis.(3) **Two novel robustness metrics** that enable users to immediately grasp how data shifts impact model performance: *Robustness Grade (RG)* that quantifies robustness when access to the training set is available or when its statistical data distribution can be estimated; and *Robustness-Aware Performance (RAP)* that measures the model's consistency under distribution shifts, focusing on its ability to maintain stable performance across different perturbations and clinical conditions.

2 Preliminaries on Robustness in Medical Imaging

Robustness of an AI system refers to its ability to consistently maintain reliable performance across diverse conditions like distribution shifts and real-world uncertainties. A robust model should generalize well beyond its training data, effectively handling previously unseen variations while avoiding catastrophic failures that compromise reliability. Medical images are subject to substantial variability. Differences in imaging devices, acquisition protocols, patient demographics, or disease prevalence can introduce distribution shifts between training and test data and, in turn, impact the robustness of AI models in medical imaging. Following Fuchs *et al.* [7], distribution shifts in medical domain can be categorized into five main types:

1. **Acquisition shift**: Variations in imaging devices, scanner settings, or manufacturers can lead to discrepancies in image quality, contrast, and resolution.
2. **Population shift**: Differences in demographic distributions (e.g., age, sex, ethnicity) between training and deployment populations can impact model generalizability.
3. **Prevalence shift**: Changes in the relative frequency of conditions across datasets can affect decision thresholds and predictive performance.
4. **Concept shift**: Evolution in clinical guidelines, annotation standards, or disease progression over time can alter the fundamental nature of segmentation or classification targets.
5. **Malicious attack**: Deliberate manipulations of data by external adversaries, such as introducing tainted samples or corrupting model updates, can compromise system integrity.

In Robust-kit, we focus on robustness to all types of data distribution shifts, excluding considerations of malicious attacks only.

3 Robust-kit

We outline here the fundamental steps and components of Robust-kit as depicted in Figure 1. Robust-kit follows a structured approach that consists of three main stages: **(1) Dataset construction,** consisting of the curation of diverse subsets that express different types of real-world data shifts; **(2) Model categorization** where different models are organized for subsequent evaluation; **(3) Robustness & performance analysis**, quantifying the severity of distribution shifts and model resilience through the newly introduced *Robustness Grade (RG)* and *Robustness-Aware Performance (RAP)* metrics.

Dataset Construction. When providing a dataset D to Robust-kit, it is necessary to indicate the type of the distribution shift S to measure and the possible sensitive attributes t for that shift. The possible sensitive attributes will vary depending on the shift but will typically include features that are relevant to fairness, domain adaptation, or subpopulation performance. These attributes help quantify the impact of distribution shifts on model behavior and ensure that robustness evaluations are contextually meaningful.

For each unique value t_i, a subset D_i is created while ensuring that other sensitive attributes remain balanced and systematically represented. This balancing serves two key purposes: (1) *Isolating the target shift* to prevent confounding factors from overshadowing the impact of the intended distribution shifts; (2) *Focused evaluation* ensuring a rigorous and unbiased assessment.To achieve this balance, we employ a stratified sampling approach, which maintains the proportional representation of sensitive attributes across subsets, ensuring robust and reliable analysis. As an example, if evaluating a dataset where *age* and *gender* are sensitive attributes, and the goal is to measure a population shift, it is crucial to ensure that *age* is balanced when assessing robustness to *gender*. An imbalance in age could introduce an unintended distribution shift, distorting the analysis.

Model Categorization. A key advantage of Robust-Kit is that it is training-free, enabling fast and efficient robustness evaluation of already trained models. We define in Robust-kit two categories of models: *Closed* and *Open* models. *Closed* models have unknown training data distributions, so robustness is assessed by evaluating performance consistency across sensitive attributes to identify potential weaknesses. *Open* models, instead, have fully or partially known training data, allowing for a direct, quantifiable robustness evaluation by measuring dataset shifts. Before running an evaluation, the user needs to categorize the model to test as either *Closed* or *Open* to ensure the appropriate robustness assessment approach is applied. **Robustness & Performance Analysis.** For both model categories, we quantitatively evaluate the impact of dataset shifts on the segmentation model's performance by first measuring the Dice Similarity Coefficient (DSC), quantifying the overlap between predicted and ground truth segmentations. Beyond standard performance evaluation, Robust-Kit introduces two novel robustness metrics: *Robustness Grade (RG)* and *Robustness-Aware Performance (RAP)*. These metrics provide a more comprehensive understanding of how models perform under distribution shifts.

For models categorized as *Open*, where the training data is known, Robust-Kit assigns a *Robustness Grade (RG)* for each distribution shift. This metric quantifies a model's ability to generalize across different shifts by incorporating three key factors: (1) Performance Retention – Evaluates how well the model maintains its Dice score between training and test datasets. (2) Severity Shift – Measures the divergence between training and test distributions using the Kullback-Leibler (KL) divergence [11]. (3) Fairness Consistency – Penalizes large variations in Dice scores across sensitive test subsets to ensure equitable predictions. These correspond to the three factors in the equation that defines RG_i. The Robustness Grade (RG_i) for each subset D_i is formally defined as:

$$RG_i = \mu \left(1 - \left|\frac{\varDelta DSC_i}{DSC_{\text{train}}}\right|\right) \frac{1}{1 + KL(D_{\text{train}}||D_i)} \left(1 - \frac{\sigma_{DSC}}{DSC_{\text{train}}}\right) \tag{1}$$

where $\varDelta DSC_i = DSC_{\text{test},i} - DSC_{\text{train}}$. Here, DSC_{train} and $DSC_{\text{test},i}$ denote the Dice scores on the training set and test subset D_i, respectively. The term $KL(D_{\text{train}}||D_i)$ quantifies the shift between the training and test distributions with the KL divergence, and σ_{DSC} denotes the standard deviation of Dice scores across test subsets. We set $\mu = 5$ so the RG_i score ranges from 0 to 5, with higher values indicating greater robustness to distribution shifts.

Table 1. Overview of datasets with associate data types, measurable shifts, and sensitive attributes.

Dataset	**Data Type**	**Measurable Shift and Sensitive Attributes**
BRATS Africa [1]	MRI	Population (Ethnicity)
AMOS [10]	CT	Population, Acquisition (Sex, Age, Manufacturer)
MnMs2 [3]	MRI	Population, Acquisition (Age, Sex, Vendor, Centre)
TOTALSEG MRI (TS MRI) [2]	MRI	Population, Acquisition (Age, Gender, Manufacturer, Scanner Model, Institute)
TOTALSEG CT (TS CT) [23]	CT	Population, Acquisition (Age, Gender, Manufacturer, Scanner Model, Institute)
OASIS2 [17]	MRI	Population, Prevalence, Concept (Age, Sex, Group, Temporal data)

For *Closed* models, we propose a modification of the Fairness-Aware Performance (FAP) metric introduced by Wu *et al.* [24] that was designed to quantitatively assess the effectiveness of multimodal large language models (MLLMs) in medical tasks. In our Robust-kit, we introduce a refined version that focuses on evaluating the performance consistency of the model as an *expression* of robustness. This approach discourages models from excelling in certain subgroups while underperforming in others, thereby promoting more equitable performance across all groups. This focus on impartial performance aligns with the ethical considerations discussed in [20] by Sabuncu *et al.*, where performance is paramount. We define this modified variance-based metric as the *Robustness-Aware Performance (RAP)*:

$$RAP = \lambda \left(\frac{\sum_{i=1}^{N} W_i \cdot DSC_{test,i}}{\sum_{i=1}^{N} W_i} - \sqrt{\frac{\sum_{i=1}^{N} W_i \cdot (DSC_{test,i} - D\bar{S}C)^2}{\sum_{i=1}^{N} W_i}} \right) \tag{2}$$

where $DSC_{test,i}$ is the Dice score for each test subset and W_i is their weight. In our implementation, we set $W_i = 1$ when no previous performance of the model is available. The first term of the equation is the weighted average Dice score, and the second term is the weighted standard deviation of the Dice score. We set $\lambda = 5$ so the RAP score ranges from 0 to 5 to be consistent with RG_i. The higher RAP, the more consistent are the performances of the model under the data shit.

4 Experiments

We present the experimental setup used to validate Robust-Kit, detailing the datasets and models employed. The primary objective of these experiments is

Table 2. Models tested for different data shifts across various datasets. The last column lists the datasets on which each model was evaluated.

Model	Acquisition	Population	Prevalence	Concept	Datasets
SegResNet [22,19,18]	✓	✓	✗	✗	AMOS, BRATS-africa
SwinUnetR [8]	✓	✓	✗	✗	AMOS
ResidualUnet [12]	✓	✓	✗	✗	MnMs2
MedSAM [15]	✓	✓	✓	✓	All
SAM-Med2D [5]	✓	✓	✓	✓	All

twofold: first, to demonstrate that our framework effectively highlights performance variations in models when exposed to distributional shifts, and second, to illustrate the practical utility of our newly proposed metrics in providing a quantifiable measure of model robustness.

Datasets. Table 1 summarizes the datasets selected for validating the Robust-Kit methodology. The selection process was guided by the need to evaluate robustness across a wide range of distributional shifts, emphasizing diversity in imaging modalities, anatomical structures, and data sources. One of the primary challenges during dataset selection was the limited availability of metadata related to patient demographics and clinical attributes. Since Robust-Kit aims to systematically analyze distribution shifts in a controlled environment, access to such metadata is crucial for accurate robustness assessment.

Models. We use Foundation Models (FMs) as *Closed* models, specifically, MedSAM [15] and SAM-Med2D [5], and we systematically assess their performance consistency using the *RAP* metric 2. For *Open* models, we consider MONAI models [4], for which we obtain a quantitative assessment through the *RG* metric 1. Table 2 provides a summary of the models evaluated across different datasets under various distribution shifts.

5 Results

Figure 2 presents a selection of results from our Robust-Kit framework. Due to space constraints, only a subset is shown here, while the full results will be available in our GitHub repository. For each shift, model, and sensitive value, we report the Dice score, while for *open* models, the *Robustness Grade (RG)* is also provided. Additionally, the pretrained models' baseline performance is indicated by a dotted line when available.

The results reveal significant variations in Dice scores across sensitive values within each dataset for some models. This variability is particularly evident in the acquisition shift of Residual U-Net and SAM-Med2D on the MnMs2 dataset. A striking example is the performance drop of SwinUnetR on the GE manufacturer subset, where the Dice score is substantially lower than on other manufacturers, suggesting poor generalization. In this case, the corresponding robustness

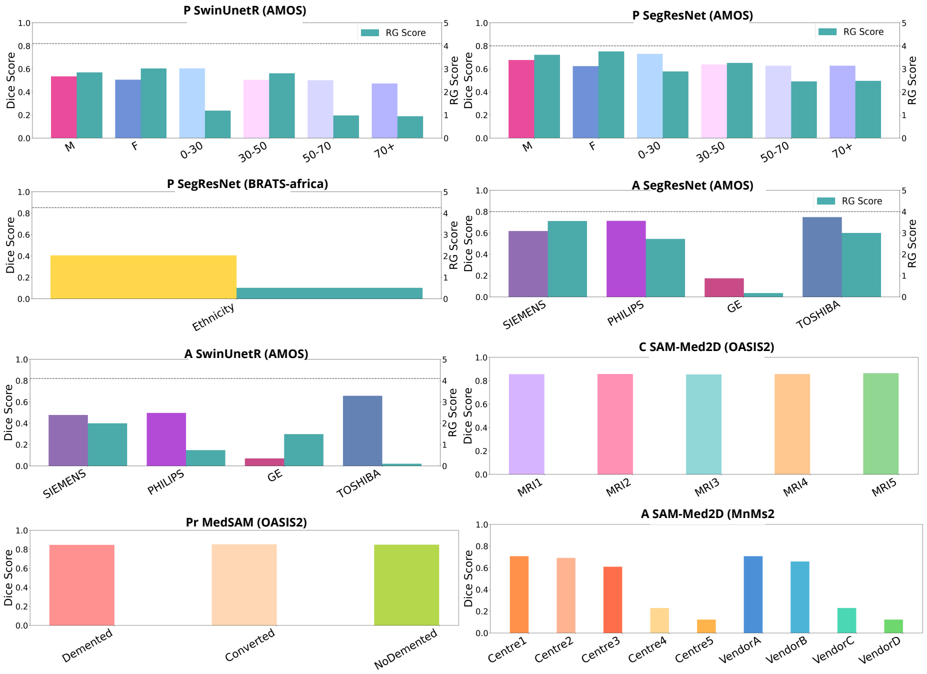

Fig. 2. A selection of results from Robust-Kit, showcasing some relevant outcomes. Each graph represents a specific distribution shift, indicated in the title (e.g., P = population shift, A = acquisition shift, C = concept shift, Pr = prevalence shift), along with the corresponding model and dataset. For open models, the *RG* metric is highlighted in green and ranges from 0 to 5. For each shift, we report the Dice score per sensitive value, consistently represented by distinct colors for improved readability.

grade (RG_{GE}) is approximately 1, illustrating the effectiveness of the RG metric in quantifying robustness. The overall robustness grade ($TotalRG = 3$) suggests that while SwinUnetR achieves strong performance in certain cases, its stability across the dataset is limited, raising concerns about its reliability in clinical applications.

In contrast, the MedSAM model demonstrates strong robustness, maintaining consistent performance across all subsets, as confirmed by the *Robustness-Aware Performance (RAP)* measure. Figure 3 shows that MedSAM consistently achieves high RAP scores across different shifts, indicating its lower susceptibility to domain shifts and making it a reliable choice for real-world medical segmentation tasks. Finally, the lowest RAP score in our study is observed for SAM-Med2D under the acquisition shift, further highlighting its sensitivity to distribution changes. As depicted in Figure 2, the Dice score for this scenario fluctuates significantly across subsets, reinforcing the model's instability under acquisition shifts. Conversely, MedSAM remains more stable, further confirming its robustness.

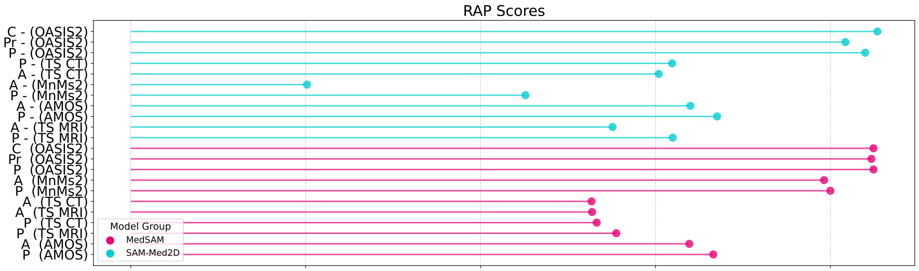

Fig. 3. Each lollipop represents the Robustness-Aware Performance (RPA) value achieved by a model under a specific distribution shift, indicated on the left (e.g., P = population shift, A = acquisition shift, C = concept shift, Pr = prevalence shift), alongside the corresponding dataset. We differentiate between MedSAM RPAs (pink) and SAM-Med2D RPAs (turquoise) for clear comparison.

6 Conclusion

In this paper, we introduced Robust-Kit, a framework designed to evaluate model robustness without requiring retraining on specific datasets to assess performance degradation. By integrating two novel metrics, *Robustness Grade (RG)* and *Robustness Assessment Percentage (RAP)*, Robust-Kit enables a comprehensive robustness assessment for both models with explicitly known training data and those where the training set is not available.

Our experimental results validate the effectiveness of Robust-Kit in quantifying robustness under various distribution shifts. By leveraging the Dice score alongside the proposed metrics, the framework provides a systematic and insightful robustness analysis, allowing users to identify vulnerabilities and performance trends across different conditions.

While our study demonstrates the utility of Robust-Kit through specific models and datasets, its applicability extends beyond these cases. The framework is highly adaptable, enabling users to integrate any dataset or model of interest for evaluation. By following the Robust-kit three-step methodology, users can efficiently and systematically assess model robustness in a standardized manner. This adaptability makes Robust-Kit a significant step toward establishing a standardized approach for robustness evaluation in medical imaging, facilitating the development and selection of more reliable models for real-world clinical applications. Future work will focus on expanding the framework's capabilities, incorporating additional critical scenarios, such as adversarial perturbations, to further enhance robustness assessment.

References

1. Adewole, M., Rudie, J., Gbadamosi, A., Zhang, D., Raymond, C., Ajigbotoshso, J., Toyobo, O., Aguh, K., Omidiji, O., Akinola, R., Suwaid, M., Emegoakor, A.,

Ojo, N., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Onuwaje, M., Betiku, B., Saluja, R., Menze, B., Baid, U., Bakas, S., Dako, F., Fatade, A., Anazodo, U.: Expanding the brain tumor segmentation (brats) data to include african populations (brats-africa) (version 1). [Dataset] (2024). https://doi.org/10.7937/v8h6-8ÃŮ67

2. Akinci D'Antonoli, T., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiss, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., Walter, A., Merkle, E.M., Boll, D.T., Breit, H.C., Nicoli, A.P., Segeroth, M., Cyriac, J., Yang, S., Wasserthal, J.: Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. Radiology **314**(2) (Feb 2025). https://doi.org/10.1148/radiol.241613, http://dx.doi.org/10.1148/radiol.241613
3. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
4. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: Monai: An open-source framework for deep learning in healthcare (2022), https://arxiv.org/abs/2211.02701
5. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023), https://arxiv.org/abs/2308.16184
6. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
7. Fuchs, M., Angelopoulos, A.N., Paschali, M., Baumgartner, C., Mukhopadhyay, A.: Navigating the unknown: out-of-distribution detection for medical imaging. In: Trustworthy AI in Medical Imaging, pp. 73–99. Elsevier (2025)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images (2022), https://arxiv.org/abs/2201.01266
9. Javed, H., El-Sappagh, S., Abuhmed, T.: Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications. Artificial Intelligence Review **58**(1), 1–107 (2025)
10. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
11. Joyce, J.M.: Kullback-leibler divergence. International encyclopedia of statistical science pp. 720–722 (2011)
12. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A.: Left-ventricle quantification using residual u-net. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9. pp. 371–380. Springer (2019)

13. Kuş, Z., Aydin, M.: Medsegbench: A comprehensive benchmark for medical image segmentation in diverse data modalities. Scientific Data **11**(1), 1283 (2024)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M.B., Gales, M.J., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.J., et al.: Shifts 2.0: Extending the dataset of real distributional shifts. arXiv preprint arXiv:2206.15407 (2022)
17. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. Journal of cognitive neuroscience **22**(12), 2677–2684 (2010)
18. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization (2018), `https://arxiv.org/abs/1810.11654`
19. Myronenko, A., Siddiquee, M.M.R., Yang, D., He, Y., Xu, D.: Automated head and neck tumor segmentation from 3d pet/ct hecktor 2022 challenge report. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 31–37. Springer (2022)
20. Sabuncu, M.R., Wang, A.Q., Nguyen, M.: Ethical use of artificial intelligence in medical diagnostics demands a focus on accuracy, not fairness (2025)
21. Schiappa, M.C., Azad, S., Vs, S., Ge, Y., Miksik, O., Rawat, Y.S., Vineet, V.: Robustness analysis on foundational segmentation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1786–1796 (2024)
22. Tang, Y., Gao, R., Lee, H.H., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M.R., Abramson, R.G., et al.: High-resolution 3d abdominal segmentation with random patch network fusion. Medical image analysis **69**, 101894 (2021)
23. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (Sep 2023). `https://doi.org/10.1148/ryai.230024`, `http://dx.doi.org/10.1148/ryai.230024`
24. Wu, P., Liu, C., Chen, C., Li, J., Bercea, C.I., Arcucci, R.: Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. arXiv preprint arXiv:2410.01089 (2024)