Silent Failures of Right-Ventricular Ejection Fraction at High Segmentation Overlap: Evidence from Multi-Vendor Cardiac MRI

Emmanouil Gionanidis[1⋆], Panagiotis Vrachnos[2], Andreas Miltiadous[1], and Alexandros T. Tzallas[1]

[1] Department of Informatics and Telecommunications, University of Ioannina, Kostakioi, Arta, 47150, Greece
[2] Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thermi-Thessaloniki, 57001, Greece
e.gionanidis@uoi.gr, panosvrachnos@iti.gr, a.miltiadous@uoi.gr, tzallas@uoi.gr

Abstract. The Dice similarity coefficient (DSC) is the default quality gate for cardiac MRI segmentation, yet clinical decisions depend not on overlap but on *derived measurements*, such as ejection fraction (EF) and ventricular volumes. We ask whether a high Dice actually guarantees that these downstream measurements are clinically reliable. Using a single frozen 2D U-Net trained on ACDC and deployed without refitting on the held-out ACDC test set (n=50) and the external multi-vendor M&Ms cohort (n=320; Siemens/Philips/GE/Canon), we define a *silent-failure rate*: among cases with conventionally high overlap (DSC ≥ 0.90), the fraction whose derived measurement exceeds a literature-derived repeatability benchmark. Right-ventricular ejection fraction (RVEF) shows a silent-failure rate of **0.52** (95% CI 0.34–0.69) on ACDC and **0.41** (0.28–0.56) on M&Ms, despite mean RV Dice of 0.887 and 0.826, respectively. The effect is observed in a pooled external cohort spanning four scanner vendors and persists across convolutional and transformer segmentation architectures. By contrast, LVEF, LVEDV, and LVSV are *protected* in-distribution, meaning that their observed silent-failure rates are near zero ($\leq$ 0.02), although LVESV is not; under shift, LV protection degrades noticeably. A mechanism analysis argues against catastrophic cancellation (end-diastolic and end-systolic errors are positively correlated) and instead identifies a systematic, same-sign volume bias to which the unsigned Dice is structurally blind, and which becomes more one-directional under shift. We conclude that a high Dice must not be used as a stand-alone safety gate for RV function, and that segmentation quality assurance should be reported against the clinical endpoint, not overlap alone.

Keywords: Cardiac MRI · Segmentation metrics · Quality assurance · Right ventricle · Distribution shift · Clinical reliability.

⋆ Corresponding author: e.gionanidis@uoi.gr

1 Introduction

Automated segmentation of the cardiac chambers from cine MRI is rarely the clinical endpoint in itself. It is an intermediate step: the segmentation yields *quantitative measurements*, i.e., ejection fraction, ventricular volumes, myocardial mass, and it is these that inform diagnosis, risk stratification, and therapy [19, 2]. Yet the dominant metric for validating and monitoring such models is the Dice similarity coefficient (DSC), the go-to gate for whether a segmentation is "good enough" to trust. This embeds a silent assumption, *that high Dice implies a clinically reliable measurement*, that is rarely tested and easy to doubt: DSC is unsigned (treating over- and under-segmentation alike) and agnostic to where errors fall, whereas a derived measurement such as EF is a *signed, ratio-valued* functional. A model can therefore achieve excellent overlap while unequal directional errors at ED and ES shift EF beyond a repeatability-derived benchmark.

Community efforts have catalogued the pitfalls of overlap metrics in the abstract [13, 20], and automated CMR EF is known to disagree across vendors and algorithms [4, 7]. Missing is a direct, quantified answer to a deployment question: *conditional on high Dice, how often does the derived measurement still fail, and for which measurements?* We provide it, both in-distribution and under realistic multi-vendor shift.

Contributions.

1. We define a *silent-failure rate*: the probability that a derived cardiac measurement exceeds a literature-derived repeatability benchmark *given* that the underlying segmentation has high Dice, and estimate it with case-level confidence intervals across eight biventricular volume and function measurements.
2. Using a single frozen model deployed on the ACDC held-out test set and the external four-vendor M&Ms cohort, we show that *right-ventricular ejection fraction silently fails in a large fraction of high-Dice cases* (0.52 on ACDC; 0.41 on M&Ms). This signature persists across four segmentation architectures spanning convolutional and transformer designs, whereas LVEF, LVEDV, and LVSV, but not LVESV, are protected in-distribution.
3. We provide a mechanism analysis that *argues against* difference amplification (catastrophic cancellation) and instead attributes the effect to a systematic same-sign volume bias invisible to the unsigned Dice, a bias that becomes more one-directional under distribution shift.

2 Related Work

Critiques of segmentation metrics. The *Metrics Reloaded* framework [13, 20] systematises how overlap metrics mislead, especially for small or thin structures, and Dice is further shown to be unreliable under uncertain, small, or empty references [18]. These establish in general terms that Dice is an imperfect proxy; we add a direct empirical quantification tied to specific clinical endpoints and

to distribution shift. Label-free quality control via reverse classification accuracy [21] predicts overlap without references, but targets DSC itself rather than the clinical endpoint which is the very quantity we show DSC fails to certify.

Overlap and volumetric accuracy. Volumetric error is closer than overlap to the clinical quantities reported from segmentations, but EF is not a single-frame volume: it is a *ratio* of EDV and ESV. High per-frame overlap can therefore coexist with clinically relevant EF error when signed ED/ES errors are correlated or systematically biased. Our contribution is to quantify this conditional failure rate directly for CMR endpoints, where the clinical question is not whether Dice is high, but whether the derived measurement remains within its benchmark.

Failure and mis-estimation detection. A growing literature distinguishes *failure detection* from OOD detection, arguing that what matters operationally is whether a prediction is wrong, not whether the input is atypical [12, 22, 23]. We share this aim but take the *derived measurement*, not pixel accuracy, as the unit of failure which behaves qualitatively differently.

Reliability of automated CMR function. Biventricular function from CMR has been benchmarked against experts and across vendors [5, 4, 7], with inter-vendor and inter-algorithm variability a known concern. The M&Ms challenge [5] and its RV-focused successor M&Ms-2 [14] target multi-vendor generalisation, with nnU-Net [11] the strongest reported solution. We use M&Ms not to benchmark a method but to test whether overlap quality *certifies* downstream RV function under the vendor shift the dataset was built to expose.

3 Materials and Methods

3.1 Model and Training

We train a single 2D U-Net (`monai.networks.nets.UNet`, MONAI 1.5.2 [6]; *not* `DynUNet`; encoder channels 32-64-128-256-512, strides 2, two residual units per stage, instance normalisation, dropout 0.1) on the ACDC training set. Input slices are resampled in-plane to 1.25×1.25 mm, centre-cropped/padded to 256×256, and z-score normalised per slice. We optimise a combined Dice–cross-entropy loss with Adam (learning rate 10^{-3}, cosine annealing) for up to 100 epochs and batch size 16. The 100 training patients are split 85/15 (*stratified by diagnostic group*; split seed 42). Validation LVEF MAE is evaluated every five epochs, and the checkpoint attaining the minimum is retained for each architecture. All retained checkpoints are frozen before test evaluation, with no fine-tuning, re-training, or re-calibration.

Three further backbones, namely SegResNet [16], Attention U-Net [17], and SwinUNETR [10], use the same data split, preprocessing, objective, optimisation schedule, and checkpoint-selection rule. SwinUNETR uses TF32 and mixed precision, whereas the convolutional models use FP32. They appear in Tables 1 and 3 and are analysed in Sec. 4.5.

3.2 Data

ACDC (in-distribution). The 50 held-out ACDC [3] test patients (IDs 101–150), spanning the five diagnostic groups, are used as the in-distribution cohort.
M&Ms (external shift). All 320 annotated M&Ms patients are used as an external cohort acquired on four scanner vendors (Siemens n=95, Philips n=125, GE n=50, Canon n=50) and spanning multiple pathologies. M&Ms uses the opposite LV/RV label convention to ACDC; predictions and references are mapped explicitly to the corresponding structures before evaluation. Volumes are computed on the common processed grid with identical voxel scaling for predictions and references, mirroring the in-distribution pipeline.
Data use. Both cohorts are publicly available for research under their respective challenge data-use agreements. ACDC and M&Ms were released by their organisers as fully anonymised data with ethical approval obtained by the original acquiring institutions; the present work is a secondary analysis of these public datasets and required no additional ethical approval.

3.3 Derived Measurements and Repeatability-Derived Benchmarks

From each segmentation we derive eight biventricular volume and function measurements per patient, end-diastolic and end-systolic volume, ejection fraction, and stroke volume for both ventricles. For each measurement we use a literature-derived repeatability benchmark: left ventricular smallest-detectable-change values from Moody et al. [15] and right ventricular interstudy-reproducibility values from Grothues et al. [9]. These are LVEF 6 EF-points, LVEDV 13 mL, and LVESV 7 mL; the RV benchmarks are relative to the reference value (RVEDV 6.2%, RVESV 14.1%, and RVEF 8.3%), with stroke-volume benchmarks propagated from their constituent volumes. In particular, $\tau_{\mathrm{RVEF},i} = 0.083|\mathrm{RVEF}_{\mathrm{ref},i}|$, not a fixed threshold of 8.3 EF-points. A measurement fails when $|m_{\mathrm{pred}} - m_{\mathrm{ref}}|$ strictly exceeds its benchmark. RVEF is prognostically important in right-heart disease [1], motivating endpoint-level evaluation. Because the LV and RV values originate from different reproducibility studies, populations, and acquisition eras, they are treated as clinical anchors rather than diagnostic decision boundaries; Sec. 4.3 also tests a common absolute EF threshold.

3.4 Silent-Failure Rate

For a measurement m we condition on conventionally high segmentation quality, $\mathrm{DSC}_{\mathrm{struct}(m)} \geq 0.90$, and define the *silent-failure rate* as the fraction of high-Dice cases for which m fails:

$$\mathrm{SF}(m) = \Pr\big[\, |m_{\mathrm{pred}} - m_{\mathrm{ref}}| > \tau_m \;\big|\; \mathrm{DSC}_{\mathrm{struct}(m)} \geq 0.90 \,\big]. \tag{1}$$

Here $\mathrm{DSC}_{\mathrm{struct}}$ is a *patient-level 3D* Dice, not a mean of per-slice values: ED and ES hard-label slices are stacked into separate volumes, Dice is computed over each complete volume, and the two phase scores are averaged with equal

weight. RVEDV, RVESV, RVEF and RVSV therefore all condition on the same RV score, and the threshold is inclusive. Slices empty in both prediction and reference contribute no voxels to either numerator or denominator. We refer to this submitted definition as the mean-phase gate. As a sensitivity analysis, we also require $\min(D_{\mathrm{ED}}, D_{\mathrm{ES}}) \geq 0.90$ and report the fraction of the complete cohort retained by this phase-strict gate. We report SF with Wilson 95% CIs, valid for the small high-Dice subgroups that arise under shift, and complement it with Dice-as-detector AUROC: the ability of $(-\mathrm{DSC})$ to rank measurement failures.

3.5 Mechanism Analysis

To probe mechanism, we report ED–ES volume-error correlation, signed RVESV error, and its signed-to-absolute ratio, $\overline{\Delta\mathrm{RVESV}}/|\overline{\Delta\mathrm{RVESV}}|$, and decompose signed RVEF error exactly into EDV and ESV contributions. For reference EDV and ESV D, S and their signed prediction errors δ_D, δ_S, the contributions are $100S\delta_D/[D(D+\delta_D)]$ and $-100\delta_S/(D+\delta_D)$; their sum is the patient-level signed RVEF error.

4 Results

4.1 Segmentation Accuracy and Derived-Measurement Error

Table 1 reports, per architecture, RV cavity Dice and the per-patient LVEF and RVEF error as median [IQR], a summary robust to the occasional degenerate segmentation that inflates a mean. Segmentation quality is moderate-to-high and stable across architectures (mean RV Dice 0.79–0.89), yet the median RVEF error is large: for the baseline, roughly 5 EF-points in-distribution and 7 under shift. RVEF error consistently exceeds LVEF error in every model, at least twofold in-distribution and 1.2–1.7× under shift.

4.2 Silent-Failure Rates Across Measurements

Table 2 reports the silent-failure rate per measurement; the RVEF pattern is consistent across cohorts. Under the endpoint-specific benchmarks, **RVEF is the worst-protected measurement**: among high-Dice cases it still fails in 14/27 cases on ACDC (0.52, CI 0.34–0.69) and 18/44 on M&Ms (0.41, 0.28–0.56), and the effect remains elevated in the pooled external four-vendor cohort. **LVEF, LVEDV, and LVSV are protected in-distribution**, here meaning observed rates ≤ 0.02 on ACDC, whereas LVESV remains an exception at 0.19 [0.10, 0.33]. Under shift, **LV protection degrades across the board**: every LV endpoint shows an increased silent-failure rate on M&Ms.

Consistent with this, Dice-as-detector AUROC for ranking measurement failures is strong for LVEF (0.979 ACDC, 0.794 M&Ms) but weak for RVEF (0.677, 0.650): Dice therefore ranks measurement failures less effectively for RVEF than for LVEF in these cohorts.

Table 1. Per-architecture RV cavity Dice and derived-measurement error for the in-distribution ACDC test set (n=50) and the external multi-vendor M&Ms cohort (n=320). The RV Dice column reports mean per-patient RV cavity Dice; the LVEF and RVEF columns report the median [interquartile range] of the per-patient absolute EF error in EF-points, robust to a small number of degenerate segmentations that inflate a mean. Moderate-to-high, stable RV Dice coexists with clinically large RVEF error in every architecture, and RVEF error consistently exceeds LVEF error.

Cohort	Architecture	RV Dice	LVEF	RVEF
ACDC	2D U-Net	0.887	2.1 [0.8, 3.7]	5.3 [2.8, 7.4]
	SegResNet	0.847	1.8 [0.7, 3.4]	5.5 [2.1, 9.7]
	Attention U-Net	0.889	1.2 [0.5, 3.4]	4.2 [2.2, 8.2]
	SwinUNETR	0.847	3.0 [1.2, 6.2]	6.4 [2.3, 9.7]
M&Ms	2D U-Net	0.826	4.2 [2.0, 7.2]	7.2 [3.8, 12.5]
	SegResNet	0.789	4.2 [1.8, 8.0]	7.0 [3.4, 13.3]
	Attention U-Net	0.844	4.3 [2.1, 7.3]	5.1 [2.3, 9.5]
	SwinUNETR	0.792	4.7 [2.6, 7.8]	6.8 [2.8, 11.9]

Table 2. Silent-failure rates under the submitted mean-phase Dice gate, with Wilson 95% CIs. Count columns report failures/accepted cases.

	ACDC (n=50)		M&Ms (n=320)	
Measurement	Failures/n	SF [95% CI]	Failures/n	SF [95% CI]
RVEF	14/27	**0.52 [0.34, 0.69]**	18/44	**0.41 [0.28, 0.56]**
RV SV	8/27	0.30 [0.16, 0.48]	9/44	0.20 [0.11, 0.35]
RVEDV	5/27	0.19 [0.08, 0.37]	16/44	0.36 [0.24, 0.51]
RVESV	4/27	0.15 [0.06, 0.32]	13/44	0.30 [0.18, 0.44]
LVESV	8/42	0.19 [0.10, 0.33]	45/141	0.32 [0.25, 0.40]
LVEF	1/42	0.02 [0.00, 0.12]	19/141	0.13 [0.09, 0.20]
LVEDV	0/42	0.00 [0.00, 0.08]	33/141	0.23 [0.17, 0.31]
LV SV	0/42	0.00 [0.00, 0.08]	26/141	0.18 [0.13, 0.26]

4.3 Sensitivity Analyses

Scaling the endpoint-specific EF benchmarks from 0.5× to 1.5× preserves the higher RVEF rate in both cohorts. Requiring both ED and ES Dice to reach 0.90 reduces the U-Net result from 14/27 to 8/17 on ACDC and from 18/44 to 6/21 on M&Ms, but coverage falls to 17/50 and 21/320. Under a common 6-EF-point threshold among patients passing both ventricular gates, RV/LV failures are 5/22 versus 0/22 on ACDC but 7/32 versus 7/32 on M&Ms. Thus, the high-Dice RVEF vulnerability is robust, whereas a universal external RV/LV ordering is not.

4.4 Per-Vendor Subgroup Analysis

A natural question is whether RVEF silent failure varies across external vendor subgroups. Conditioning on high Dice *and* splitting by vendor leaves small subgroups (high-Dice RVEF counts 6, 25, 2, and 11 for Siemens, Philips, GE, and Canon), with elevated but wide, overlapping rates: 4/6 (0.67), 9/25 (0.36), 1/2 (0.50), and 4/11 (0.36). We therefore do not claim a vendor gradient: the pooled external estimate remains elevated, but these data cannot resolve a vendor ordering.

4.5 Robustness Across Architectures

To assess whether the finding is specific to the baseline U-Net, we repeat the submitted mean-phase analysis for the three additional backbones. RVEF silent failures recur in every architecture and cohort (Table 3). Because the models share data, preprocessing, and evaluation, and accept different patient subsets, this is a shared-protocol robustness check rather than independent replication or a direct model ranking.

Table 3. RVEF silent-failure rates under the submitted mean-phase Dice gate across four 2D backbones, with Wilson 95% CIs. Count columns report failures/accepted cases.

	ACDC (n=50)		M&Ms (n=320)	
Architecture	Failures/n	RVEF SF [95% CI]	Failures/n	RVEF SF [95% CI]
2D U-Net	14/27	0.52 [0.34, 0.69]	18/44	0.41 [0.28, 0.56]
SegResNet	8/14	0.57 [0.33, 0.79]	13/33	0.39 [0.25, 0.56]
Attention U-Net	11/27	0.41 [0.25, 0.59]	21/79	0.27 [0.18, 0.37]
SwinUNETR	7/17	0.41 [0.22, 0.64]	14/42	0.33 [0.21, 0.48]

4.6 Mechanism: Signed Bias versus Difference Amplification

ED and ES RV volume errors are positively correlated (0.39 ACDC; 0.70 M&Ms), so independent, oppositely signed phase errors are not the dominant explanation. The dominant signed bias is positive RVESV error: mean signed RVESV error is +3.79 mL on ACDC and +7.45 mL on M&Ms, positive in 19/27 and 36/44 high-Dice cases, respectively (Fig. 1b). The signed-to-absolute error ratio increases from 0.45 on ACDC to 0.86 on M&Ms, consistent with a more one-directional bias under shift. Exact decomposition localises this further: mean EDV contributions are $+0.52/+1.25$ EF-points on ACDC/M&Ms, whereas mean ESV contributions are $-2.43/-3.95$, yielding mean signed RVEF errors of $-1.91/-2.70$ EF-points. The signed volume-bias explanation is therefore refined as an ESV-dominated cross-phase imbalance: Dice encodes neither error direction nor the relative ED/ES balance.

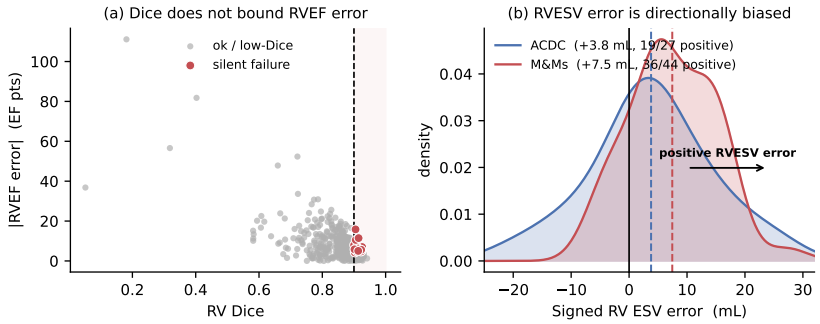

Fig. 1. Mechanism. (a) On M&Ms, high RV Dice does not bound RVEF error: cases passing the 0.90 gate can still exceed the RVEF benchmark. (b) Signed RVESV error within the high-Dice subgroup; positive errors predominate (19/27 ACDC; 36/44 M&Ms), consistent with the ESV-dominated negative RVEF contribution quantified in Sec. 4.6.

5 Discussion

Our central finding is narrow but practically important: *a high Dice does not certify right-ventricular ejection fraction*. In a substantial fraction of high-Dice cases, RVEF lies outside its repeatability-derived benchmark, in-distribution, in a pooled external four-vendor cohort, and across four architectures including a transformer. The mechanism is mundane but dangerous: a systematic, directional volume bias that an unsigned overlap metric cannot see, and that becomes more one-directional under distribution shift. The phase-strict sensitivity and exact decomposition refine rather than alter this interpretation: failures remain when both ED and ES Dice exceed 0.90, and the ESV contribution dominates the signed RVEF error.

Implications. Gate on the *clinical endpoint*, not Dice alone. Where reference measurements are unavailable at deployment, our results caution rather than solve: a useful monitor must be sensitive to directional endpoint error, not merely to low overlap or generic uncertainty (e.g. MC-dropout [8]). This points to a directional-bias-aware monitor rather than a generic confidence score.

Exploratory stratification by diagnosis showed particularly large positive signed RVESV error in DCM-labelled high-Dice cases (ACDC: +11.0 mL, n=8; M&Ms: +8.3 mL, n=17). Point estimates varied in magnitude and, in some small diagnostic subgroups, in sign, suggesting that a single pooled bias correction may be insufficient.

Limitations. Absolute rates are model- and benchmark-specific. The four backbones share training data, preprocessing, checkpoint selection, and evaluation, so the architecture analysis is not independent replication and does not include controlled seed uncertainty. All models are 2D, and both cohorts are retrospective public datasets. Finally, strict gates produce small accepted subsets, and

the different LV/RV benchmark constructions limit claims of a universal cross-ventricular ordering.

6 Conclusion

Overlap accuracy and clinical reliability are not the same thing, and for the right ventricle they can come apart sharply. Across an in-distribution cohort, a pooled external cohort spanning four vendors, and four segmentation architectures, RVEF exceeded its repeatability-derived benchmark in a substantial fraction of high-Dice cases. The effect is associated with a signed RV volume bias to which Dice is blind. Quality assurance for cardiac function should therefore be expressed against the derived clinical measurement, not overlap alone.

Acknowledgements. This work was supported by the European Union's EU4Health Programme under Grant Agreement No. 101218966 (UNICA: Unified Network for International Cancer Advancement).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alabed, S., Shahin, Y., Garg, P., Alandejani, F., Johns, C.S., et al.: Cardiac-MRI predicts clinical worsening and mortality in pulmonary arterial hypertension: a systematic review and meta-analysis. JACC Cardiovasc. Imaging **14**(5), 931–942 (2021). doi:10.1016/j.jcmg.2020.08.013
2. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J. Cardiovasc. Magn. Reson. **20**, 65 (2018). doi:10.1186/s12968-018-0471-x
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). doi:10.1109/TMI.2018.2837502
4. Bhuva, A.N., Bai, W., Lau, C., Davies, R.H., et al.: A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. Circ. Cardiovasc. Imaging **12**(10), e009214 (2019). doi:10.1161/CIRCIMAGING.119.009214
5. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021). doi:10.1109/TMI.2021.3090082
6. Cardoso, M.J., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv:2211.02701 (2022)
7. Davies, R.H., Augusto, J.B., Bhuva, A., Xue, H., et al.: Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. J. Cardiovasc. Magn. Reson. **24**(1), 16 (2022). doi:10.1186/s12968-022-00846-4

8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML. PMLR, vol. 48, pp. 1050–1059 (2016)
9. Grothues, F., Moon, J.C., Bellenger, N.G., Smith, G.S., Klein, H.U., Pennell, D.J.: Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance. Am. Heart J. **147**(2), 218–223 (2004). doi:10.1016/j.ahj.2003.10.005
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion Workshop (MICCAI), LNCS, vol. 12962, pp. 272–284. Springer, Cham (2022). arXiv:2201.01266. doi:10.1007/978-3-031-08999-2_22
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021). doi:10.1038/s41592-020-01008-z
12. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. In: ICLR (2023). arXiv:2211.15259
13. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., et al.: Metrics Reloaded: recommendations for image analysis validation. Nat. Methods **21**(2), 195–212 (2024). doi:10.1038/s41592-023-02151-z
14. Martín-Isla, C., Campello, V.M., Izquierdo, C., Kushibar, K., et al.: Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge. IEEE J. Biomed. Health Inform. **27**(7), 3302–3313 (2023). doi:10.1109/JBHI.2023.3267857
15. Moody, W.E., Edwards, N.C., Chue, C.D., Taylor, R.J., Ferro, C.J., Townend, J.N., Steeds, R.P.: Variability in cardiac MR measurement of left ventricular ejection fraction, volumes and mass in healthy adults: defining a significant change at 1 year. Br. J. Radiol. **88**(1049), 20140831 (2015). doi:10.1259/bjr.20140831
16. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion Workshop (MICCAI), LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). doi:10.1007/978-3-030-11726-9_28
17. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., et al.: Attention U-Net: learning where to look for the pancreas. In: Medical Imaging with Deep Learning (MIDL) (2018). arXiv:1804.03999
18. Ostmeier, S., Axelrod, B., Isensee, F., et al.: USE-Evaluator: performance metrics for medical image segmentation models supervised by uncertain, small or empty reference annotations in neuroimaging. Med. Image Anal. **90**, 102927 (2023). doi:10.1016/j.media.2023.102927
19. Petersen, S.E., Aung, N., Sanghvi, M.M., et al.: Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. J. Cardiovasc. Magn. Reson. **19**, 18 (2017). doi:10.1186/s12968-017-0327-9
20. Reinke, A., Tizabi, M.D., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nat. Methods **21**(2), 182–194 (2024). doi:10.1038/s41592-023-02150-0
21. Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., et al.: Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. J. Cardiovasc. Magn. Reson. **21**(1), 18 (2019). doi:10.1186/s12968-019-0523-x

22. Xia, G., Bouganis, C.-S.: Augmenting softmax information for selective classification with out-of-distribution data. In: Computer Vision – ACCV 2022, LNCS, vol. 13846, pp. 664–680. Springer, Cham (2023). doi:10.1007/978-3-031-26351-4_40
23. Zhu, F., Cheng, Z., Zhang, X.-Y., Liu, C.-L.: Rethinking confidence calibration for failure prediction. In: Computer Vision – ECCV 2022, LNCS, vol. 13685, pp. 518–536. Springer, Cham (2022). doi:10.1007/978-3-031-19806-9_30