Journal of Cardiovascular Magnetic Resonance 28 (2026) 102677

Contents lists available at ScienceDirect

Journal of Cardiovascular Magnetic Resonance

journal homepage: www.sciencedirect.com/journal/jocmr

Original Research

Cardiac function assessment with deep-learning-based automatic segmentation of free-running four-dimensional whole-heart cardiovascular magnetic resonance

Augustin C. Ogier[a,*,1], Salomé Baup[a,2], Gorun Ilanjian[a], Aisha Touray[a], Angela Rocca[b,3], Jaume Banús[a,4], Isabel Montón Quesada[a], Martin Nicoletti[a], Jean-Baptiste Ledoux[a,c,5], Jonas Richiardi[a,c,6], Robert J. Holtackers[a,d,7], Jérôme Yerly[a,c,8], Matthias Stuber[a,c,9], Roger Hullin[b,10], David Rotzinger[a,11], Ruud B. van Heeswijk[a,12]

[a] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[b] Cardiovascular Department, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[c] CIBM Center for Biomedical Imaging, Lausanne, Switzerland
[d] Department of Radiology & Nuclear Medicine, Maastricht University Medical Centre, the Netherlands

ARTICLE INFO

Keywords:
Cardiac MRI
Free-running
Deep learning
Cardiac segmentation
3D + t imaging
Isotropic imaging

ABSTRACT

Background: Free-running (FR) cardiac magnetic resonance imaging (MRI) enables free-breathing electrocardiogram (ECG)-free fully dynamic five-dimensional (5D) (three-dimensional [3D] spatial + cardiac + respiration dimensions) imaging but poses significant challenges for clinical integration due to the volume of data and complexity of image analysis. Existing segmentation methods are tailored to two-dimensional (2D) cine or static 3D acquisitions and cannot leverage the unique spatial-temporal wealth of FR data. The aim of this study was to develop and validate a deep learning (DL)-based segmentation framework for isotropic 3D + cardiac cycle FR cardiac MRI that enables accurate, fast, and clinically meaningful anatomical and functional analysis.

Methods: Free-running, contrast-free balanced steady-state free precession (bSSFP) acquisitions at 1.5T and contrast-enhanced gradient-recalled echo (GRE) acquisitions at 3T were used to reconstruct motion-resolved 5D datasets. From these, the end-expiratory respiratory phase was retained to yield fully isotropic four-dimensional

Abbreviations: 2D, two-dimensional; 3D, three-dimensional; 4D, four-dimensional; ACDC, Automated Cardiac Diagnosis Challenge; ASSD, average symmetric surface distance; bSSFP, balanced steady-state free precession; CS, compressed sensing; CMR, cardiovascular magnetic resonance; DL, deep learning; DSC, Dice similarity coefficient; ED, end-diastole; EDV, end-diastolic volume; EF, ejection fraction; ES, end-systole; ESV, end-systolic volume; GRE, gradient-recalled echo; HD, Hausdorff distance; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; ICC, intra-class correlation coefficient; MMWHS, multi-modality whole-heart segmentation; PCA, principal-component analysis; RV, right ventricle/ventricular; RVD, relative volume difference; SAX, short-axis; SCMR, Society for Cardiovascular Magnetic Resonance; SEM, standard error of measurement; TR, repetition time; TE, echo time; LOA, limits of agreement; FR, free running; 5D, five-dimensional; ECG, electrocardiogram; L/RVB, left/right ventricular blood pool; LVM, left ventricular myocardiumleft ventricular myocardium; LV, left ventricle/ventricular; GPU, graphics processing unit; CPU, central processing unit; ANOVA, analysis of variance

* Corresponding author.
E-mail address: augustin.ogier@gmail.com (A.C. Ogier).
[1] ORCID: 0000–0001-9178–9964
[2] ORCID: 0009–0004-0253–5153
[3] ORCID: 0000–0002-1634–6992
[4] ORCID: 0000–0001-9318–6323
[5] ORCID: 0000–0003-0447–5073
[6] ORCID: 0000–0002-6975–5634
[7] ORCID: 0000–0003-1809–313X
[8] ORCID: 0000–0003-4347–8613
[9] ORCID: 0000–0001-9843–2028
[10] ORCID: 0000–0002-5738–1903
[11] ORCID: 0000–0002-6321–3180
[12] ORCID: 0000–0001-5028–4521

https://doi.org/10.1016/j.jocmr.2025.102677
Received 18 July 2025; Received in revised form 5 December 2025; Accepted 22 December 2025

(4D) datasets. Automatic propagation of a limited set of manual segmentations was used to segment the left and right ventricular blood pool (LVB, RVB) and left-ventricular myocardium (LVM) on reformatted short-axis (SAX) end-systolic (ES) and end-diastolic (ED) images. These were used to train a 3D nnU-Net model. Validation was performed using geometric metrics (Dice similarity coefficient [DSC], relative volume difference [RVD]), clinical metrics (ED and ES volumes, ejection fraction [EF]), and physiological consistency metrics (systole–diastole LVM volume mismatch and LV–RV stroke volume agreement). To assess the robustness and flexibility of the approach, we evaluated multiple additional DL training configurations, such as using 4D propagation-based data augmentation to incorporate all cardiac phases into training.

Results: The main proposed method achieved automatic segmentation within a minute, delivering high geometric accuracy and consistency (DSC: 0.94 ± 0.01 [LVB], 0.86 ± 0.02 [LVM], 0.92 ± 0.01 [RVB]; RVD: 2.7%, 5.8%, 4.5%). Clinical LV metrics showed excellent agreement (ICC > 0.98 for EDV/ESV/EF, bias < 2 mL for EDV/ESV, < 1% for EF), while RV metrics remained clinically reliable (ICC > 0.93 for EDV/ESV/EF, bias < 1 mL for EDV/ESV, < 1% for EF) but exhibited wider limits of agreement. Training on all cardiac phases improved temporal coherence, reducing LVM volume mismatch from 4.0% to 2.6%.

Conclusion: This study validates a DL-based method for fast and accurate segmentation of whole-heart free-running 4D cardiac MRI. Robust performance across diverse protocols and evaluation with complementary metrics that match state-of-the-art benchmarks supports its integration into clinical and research workflows, helping to overcome a key barrier to the broader adoption of free-running imaging.

1. Introduction

Cardiovascular disease remains the leading cause of death globally [1], highlighting the critical need for accurate, efficient, and reproducible cardiac imaging tools to improve diagnosis. Cardiovascular magnetic resonance imaging (CMR) plays an important role in the assessment of cardiac anatomy and function, combining high soft-tissue contrast with a wide range of diagnostic parameters [2]. However, conventional functional CMR protocols, particularly 2D multi-slice cine imaging, have limitations. These acquisitions usually require breath holding and are typically performed using thick slices (often 6–10 mm) and inter-slice gaps. Though they offer temporal coverage of the cardiac cycle, the poor through-plane resolution and dependence on predefined imaging planes may reduce their accuracy and reproducibility, especially in complex anatomies or when additional anatomical or functional information is requested after the scan has finished. Static 3D CMR acquisitions, such as whole-heart balanced steady-state free precession (bSSFP), overcome the limitations of slice thickness and provide a high (near-)isotropic spatial resolution, but they are generally acquired in a single cardiac phase (usually end-diastole) and are not meant to capture cardiac dynamics. Both 2D cine and static 3D imaging depend on ECG triggering, making them vulnerable to poor ECG signal quality, and the additional breath-hold requirements of 2D cine further challenge patients with limited breath-hold capacity. This dependence on patient cooperation and accurate timing adds complexity to the workflow and can impact image quality and reproducibility.

Recent free-breathing CMR strategies, including full free-breathing clinical protocols [3] and deep-learning–accelerated 4D cine reconstructions [4], simplify scan prescription but still rely on ECG triggering, which remains sensitive to gradient-induced interference and magnetohydrodynamic effects, particularly at higher field strengths. Multitasking CMR [5], while ECG-free through self-gating, is still implemented as stacks of 2D slices and does not provide isotropic whole-heart coverage. Free-running (FR) CMR has recently emerged as a powerful alternative for functional and anatomical cardiac imaging, offering motion-resolved, high-resolution 3D imaging of the entire heart across both cardiac and respiratory cycles [6]. This technique combines a continuous free-breathing ECG-free acquisition with self-gated motion-resolved reconstruction, resulting in fully isotropic 5D datasets (isotropic 3D spatial + cardiac + respiratory dimensions). FR imaging enables retrospective reformatting in any orientation and provides both anatomical detail and functional coverage across all phases of the cardiac cycle. However, the ease with which FR datasets can be acquired stands in sharp contrast to the complexity of analyzing the vast number of images they generate. A single dataset can contain up to 70 short-axis slices across 4 respiratory and 25 cardiac phases—amounting to around 7000 images per subject. This extensive information, while potentially clinically valuable, makes full manual segmentation unfeasible and currently prevents broader clinical adoption. As a result, prior validation and analyses of FR acquisitions have relied on careful reformatting of the isotropic volumes into a traditional stack of 2D short-axis (SAX) views to enable segmentation using conventional tools, [7–9] ultimately negating the benefits of isotropic dynamic whole-heart imaging.

Deep learning (DL) methods have become the standard for automated CMR segmentation [10], and have been validated across cohorts such as the ACDC [11], UK Biobank [12], and MM-WHS [13]. These benchmarks have played a central role in the development of fully automatic CMR segmentation methods, enabling systematic comparisons of architectures and serving as reference datasets for most segmentation studies in the field [11,12,14–16]. While these efforts have led to major advances in automatic segmentation performance, they are inherently tied to classical CMR data formats. Most 2D cine segmentation models are trained on multi-slice SAX stacks with non-isotropic resolution [11,12,14]. Segmentation of static 3D datasets provides high-resolution anatomical descriptions but is limited to a single cardiac phase and lacks dynamic information [13,15,16]. Among existing dynamic true 3D CMR techniques comparable to FR imaging, 4D flow MRI offers a precedent for 4D segmentation [17]. However, its emphasis on velocity encoding results in lower anatomical resolution and contrast, which limits its suitability for detailed structural segmentation.

To date, no DL method has been proposed to perform segmentation directly on the native isotropic 3D + cardiac dimension of FR CMR acquisitions [10]. Existing segmentation frameworks are not readily transferable due to differences in image contrast, resolution, and dimensionality. Moreover, there is no publicly available benchmark dataset that includes manual annotations for FR imaging in its native form.

The goal of this study was to develop and validate a DL-based automatic segmentation framework for isotropic 3D + cardiac cycle FR CMR that enables accurate, fast, and clinically meaningful anatomical and functional analysis. The model is trained on a diverse cohort of healthy volunteers and heart failure patients, spanning multiple contrasts and magnetic field strengths. To overcome the impracticality of manual annotation across such large and dynamic datasets, we leverage a previously validated segmentation propagation approach based on diffeomorphic registrations [18] to generate ground truth segmentation at end-diastole and end-systole. The proposed DL framework enables fast high-resolution, full-cycle segmentation directly in the original 3D + cardiac space, without relying on 2D reformats. Crucially, we extend the traditional geometric-only evaluation by also assessing key clinical

and physiological consistency metrics. By validating performance across both geometric and functional cardiac metrics, this work supports the broader clinical translation of 5D FR CMR and helps bridging the gap between advanced acquisition and clinical usefulness.

2. Methods

2.1. Study population and acquisition protocol

All human studies were approved by the local ethics committee (CER-VD approvals 2022–00934 and 2021–02458), and written informed consent was obtained from all participants before scanning. Thirty-five subjects were prospectively scanned on two clinical MRI systems to assess the generalizability of the segmentation framework across field strength and contrast regimes (Fig. 1C). Dataset 1 (D_1) [7] included 15 healthy volunteers scanned at 1.5T (Magnetom Sola, Siemens Healthineers, Forchheim, Germany) with a FR acquisition based on a native (without injection of a contrast agent) bSSFP readout. Dataset 2 (D_2) [19] included 20 participants, including 10 healthy controls, 5 patients with heart failure with preserved ejection fraction (HFpEF), and 5 patients with heart failure with reduced ejection fraction (HFrEF), imaged at 3T (Magnetom PrismaFit, Siemens

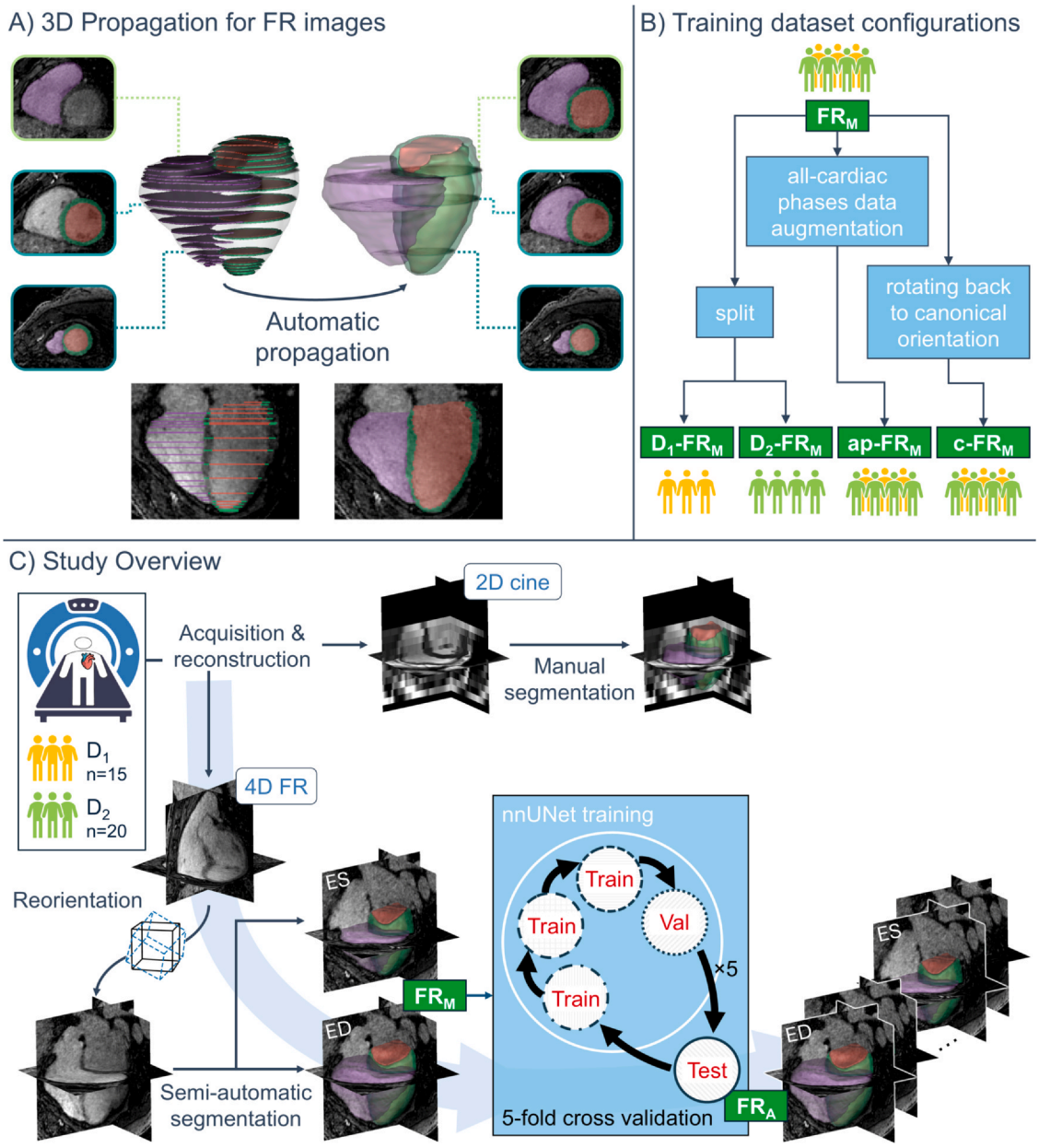

Fig. 1. Workflow for semi-automatic and deep-learning segmentations of FR CMR. A) 3D-propagation scheme: Manual contours on a limited set of slices are extended to the entire 3D volume using a combination of diffeomorphic registrations. B) Training configurations: the main configuration (FR_M) consists of semi-automatic segmentations on SAX-reformatted end-systolic (ES) and end-diastolic (ED) phases. Additional training configurations were derived to test specific hypotheses: cohort-specific training (D_1-FR_M, D_2-FR_M), 4D data augmentation including all cardiac phases (ap-FR_M), and training in the native canonical space (c-FR_M). C) Study design: Two cohorts (D_1 at 1.5 T; D_2 at 3 T) underwent 2D cine and FR imaging. FR acquisitions were reconstructed as 5D images, but only the end-expiratory bin was retained, yielding 4D datasets. Cine stacks were manually segmented, and FR volumes were reoriented and semi-automatically segmented at ED and ES via the 3D-propagation scheme in (A), yielding FR_M. The main configuration FR_M was used to train an nnU-Net model, yielding the automatic segmentation FR_A under 5-fold cross-validation. All additional training configurations in (B) followed the same workflow. *FR* free-running, *CMR* cardiovascular magnetic resonance, *3D* three-dimensional, *SAX* short-axis, *ES* end-systole, *ED* end-diastole, *4D* four-dimensional

Table 1
Participant demographics

	Dataset 1 (D_1)	Dataset 2 (D_2)		
Subjects	15 healthy	10 healthy	5 HFpEF	5 HFrEF
Age (y)	25 ± 3	69 ± 8	67 ± 10	48 ± 14
Weight (kg)	67 ± 11	73 ± 13	73 ± 8	91 ± 14
Height (cm)	172 ± 8	175 ± 10	169 ± 9	176 ± 4
Sex	7 F/8 M	5 F/5 M	1 F/4 M	0 F/5 M

HFpEF heart failure with preserved ejection fraction, *HFrEF* heart failure with reduced ejection fraction. *Data are presented as means ± standard deviation.*

Healthineers) using an FR sequence based on a spoiled gradient-recalled echo (GRE) acquired immediately after a contrast agent injection (0.2 mmol/kg gadobutrol, (Gadovist, Bayer AG, Leverkusen, Germany). The segmented 3D radial FR volumes were acquired in a canonical orientation (i.e., not oblique), centered on the left ventricle and with each consecutive radial interleave initiated by a line in the superior–inferior direction for the extraction of physiological motion signals [20]. Demographic details are provided for both cohorts (Table 1). As a reference standard, all examinations also included a 2D cine protocol that was retrospectively ECG-gated and acquired during end-expiratory breath-holds of 5–10 s. This compressed-sensing-accelerated protocol consisted of a stack of 12 SAX cine plus single-slice 4-chamber and single-slice 2-chamber cines [21]. Acquisition parameters are detailed for each pulse sequence and field strength (Table 2). All examinations were performed with the combination of a 32-channel spine coil and an 18-channel body coil.

2.2. 5D image reconstruction

For the free-running 5D images, cardiac and respiratory self-gating signals were extracted from the repeated superior–inferior projections of the segmented 3D radial trajectory [6]. First, principal-component analysis generated several candidate respiratory signals. For each candidate component, Gaussian curves were fitted to every peak within the physiological breathing band of its power spectrum [8]. A scoring metric, designed to favor components with a single dominant peak while down-weighting those with additional spurious peaks, was calculated, and the highest-scoring candidate was automatically selected as the respiratory trace [8]. After subtracting this respiratory component, second-order blind identification isolated the dominant cardiac component [22], which was selected in the expected heart-rate range using the same spectral-ranking approach.

The validated motion components were used to retrospectively sort the continuous acquisition into four equally populated respiratory bins (end-expiration to end-inspiration) and non-overlapping 50 ms cardiac frames, yielding highly undersampled 5D k-space blocks. Compressed sensing (k-t sparse SENSE) was used to reconstruct 5D cardiac and respiratory motion-resolved images [6,23], with regularization weights set to 0.001 for both temporal dimensions in D_1, and 0.01 (cardiac) and 0.03 (respiratory) in D_2. For the present study only the end-expiratory bin was retained, resulting in a 4D dataset (3D + cardiac cycle) per subject that spans the full cardiac cycle with minimal residual respiratory motion. All steps of this 5D reconstruction pipeline were fully automated and required no subjective user intervention and were performed with a reconstruction time of 261 ± 44 min per subject.

2.3. Ground truth segmentation

Segmentation followed Society for Cardiovascular Magnetic Resonance (SCMR) recommendations for biventricular analysis [24] and produced segmentation masks for the left-ventricular blood pool (LVB), left-ventricular myocardium (LVM) and right-ventricular blood pool (RVB). The annotation protocol was further informed by prior literature aimed at standardizing ventricular segmentation [25]. All contours were drawn and edited in FSLeyes [26]. Although commercial CMR analysis software packages are commonly used in clinical practice, these platforms typically do not allow the exportation of segmentation masks for subsequent DL training, so they could not be employed for this study.

For the 2D cine reference data, end-diastolic (ED) and end-systolic (ES) frames were identified by visual inspection of the LVB area. Segmentation was performed slice-by-slice on the SAX cine, while the single-slice 4-chamber cine was consulted to determine the most basal SAX slice to include. A radiologist (with 2 years of experience) manually contoured the LVB and LVM, and a second radiologist (with 2 years of experience) delineated the RVB for both ED and ES phases.

For the FR volumes, the isotropic 4D images were rotated to conventional double-oblique cardiac axes by applying a rigid transform derived from the 2D cine SAX and 4-chamber orientations, allowing the subsequent manual work to follow SCMR guidelines that are defined for

Table 2
MR image acquisition parameters for free-running (FR) and 2D cine acquisitions in datasets D_1 and D_2

Sequence	2D Cine		FR	
Dataset	D_1	D_2	D_1	D_2
Field strength	1.5T	3T	1.5T	3T
Sequence type	CS bSSFP	CS bSSFP	slab-selective bSSFP	slab-selective GRE
Contrast agent	None	None	None	0.2 mmol/kg gadobutrol
k-space trajectory	Undersampled Cartesian	Undersampled Cartesian	Segmented 3D radial spiral phyllotaxis	Segmented 3D radial spiral phyllotaxis
Dimensions	2D	2D	3D	3D
TE (ms)	1.4	1.3	1.7	1.3
TR (ms)	3.2	3.4	3.5	3.0
Flip angle (°)	80	80	Median 70 (range 58–70)	15
In-plane resolution (mm^2)	1.4 × 1.4	1.4 × 1.4	1.4 × 1.4	1.4 × 1.4
Slice thickness (mm)	8	8	1.4 — Isotropic slab	1.4 — Isotropic slab
Slice gap (mm)	2	2	0	0
Undersampling factor	7.5	8.6	2.1–4.0 Nyquist % per motion bin	1.7–4.0 Nyquist % per motion bin
Receiver bandwidth (Hz/pixel)	977	977	1116	1008
Field of view	360 × 284 mm^2	360 × 287 mm^2	220 × 220 × 220 mm^3	220 × 220 × 220 mm^3
ECG triggering	Yes	Yes	No	No
Respiratory motion handling	Breath hold	Breath hold	Self-gated	Self-gated
Acquisition time (min)	~ 10	~ 10	4:58	4:05

Both cine and FR images were reconstructed using compressed sensing. The reported cine acquisition time includes full planning (from localizers to double-oblique plane prescription) and was dependent on patient breath-hold capacity. "Nyquist % per motion bin" indicates the sampling density of each combined respiratory × cardiac motion bin, expressed as a percentage of the lines required for a fully sampled 3D radial acquisition. Data are presented as numbers and ranges, as appropriate. *2D* two-dimensional, FR free running, *3D* three-dimensional, *CS* compressed sensing, *bSSFP* balanced steady state free precession

these views. After this transformation, the main imaging plane coincided with the anatomical short-axis view, whereas the two orthogonal planes corresponded approximately to the operator-prescribed long-axis orientations. Because a single global rigid transformation cannot always simultaneously match the exact angulation of both clinical short- and long-axis cines, these orthogonal planes were thus pseudo-2-chamber and pseudo-4-chamber views. ED and ES frames were then identified by visual inspection of the LVB area. Because each phase comprised 60–70 short-axis slices, fully manual segmentation was impractical. We therefore leveraged a previously validated semi-automatic segmentation pipeline [18,27]. The same two radiologists who segmented the cine images manually contoured separate sets of key slices for ED and ES, including the most basal and apical planes, and added contours at fixed spatial steps or at levels showing significant anatomical change. After manual segmentation of the initial slices, the segmentations were automatically 3D-propagated independently for the ED and ES phases (Fig. 1A) through the remainder of the volume using a combination of diffeomorphic registration techniques [18], and the radiologists reviewed and performed minor local corrections on all propagated masks if needed. This semi-automatic workflow reduced per-subject annotation time while maintaining expert-level accuracy. The number of manually segmented and propagated slices was recorded for each subject.

Both the 2D cine and FR segmentations were subjected to two physiological consistency checks as follows: conservation of myocardial volume between ED and ES and agreement of stroke volume between the left and right ventricles. To ensure myocardial volume consistency, LVM segmentations were adjusted to maintain ED–ES differences within a small tolerance compatible with physiological expectations and imaging precision. Although several studies suggest a slight systolic decrease in myocardial volume due to transient intramyocardial blood or lymph extrusion, this effect remains below the variability introduced by partial-volume effects and contour placement in CMR. Therefore, consistent with SCMR recommendations [24] and large-cohort practices such as UK Biobank [28], myocardial volume consistency was therefore used as a quality-control measure to identify non-physiological deviations rather than to enforce strict mass conservation. LV–RV stroke volume agreement was only assessed in healthy subjects, as regurgitation in patients could confound this metric. In cases of uncertainty, RVB segmentations, most often in 2D cine, were iteratively refined to aim for an RV stroke volume matching the LV stroke volume within a 10 mL tolerance. Any segmentation that failed either check was reviewed by a senior radiologist (12 years of experience), who applied targeted adjustments until both metrics fell within clinical tolerance, yielding the definitive ground-truth masks.

Hereafter, the fully manual 2D cine segmentations will be referred to as $Cine_M$, and the semi-automatic FR segmentations as FR_M. In order to quantify whether the rotation into the double-oblique orientation is needed for the subsequent DL training, we defined a second mask set, c-FR_M (for canonical-FR_M), by applying the inverse of the cine-derived affine transform to FR_M (Fig. 1B), thereby transforming all contours into the canonical axial, sagittal, and coronal coordinate system of the original FR acquisition.

2.4. Deep learning model and training

For the DL model, we used the well-known nnU-Net [29] enhanced with residual encoder framework [30], which has proven to be very competitive in several volumetric medical image segmentation challenges and has demonstrated strong performance across diverse tasks [30]. We employed the 3D full-resolution configuration, as it provided the best results in preliminary testing. The model was trained using Dice loss, which is well-suited for medical image segmentation [31] due to its direct optimization of spatial overlap between predicted and ground-truth masks. The architecture, optimization parameters, and training schedule were fixed according to the default nnU-Net configuration [29]. The batch size was set to 4 due to graphics processing unit (GPU) memory limitations. The number of epochs was set to 80, based on empirical observation of training performance plateauing. All networks were trained on the full isotropic 3D volumes at the native field of view and resolution reported in Table 2. The automatic segmentation output consisted of four segmentation classes as follows: background, LVB, LVM, and RVB. Trainings were run on a Linux workstation with two 24-core central processing units (CPUs, Intel Xeon Gold 6248 R; Intel, Santa Clara, California), 1.5 TB of RAM, and a 48 Gb RTX A6000 GPU (NVIDIA Corporation, Santa Clara, California). Both the total training time and the inference time were recorded.

To systematically validate our model across all subjects in the main configuration (FR_M), we employed a 5-fold cross-validation strategy (Fig. 1C). The 35-subject cohort was divided into five non-overlapping folds of seven FR_M each (three D_1 and four D_2 subjects per fold). Within each fold, the D_2 subset consisted of two healthy volunteers, one HFpEF patient, and one HFrEF patient. For each iteration, 3 folds (21 subjects, 42 three-dimensional volumes corresponding to ED and ES) were used for training, 1 fold (7 subjects, 14 three-dimensional volumes) for validation, and the remaining fold (7 subjects, 14 three-dimensional volumes) for testing, ensuring each fold served exactly 3 times as training, once as validation, and once as testing.

To evaluate the impact of data origin and data augmentation on segmentation performance, we implemented three additional training dataset configurations. These followed the same 5-fold logic but with adapted splits as needed (Fig. 1B). First, we trained one model exclusively on D_1-FR_M subjects and another exclusively on D_2-FR_M subjects, thereby isolating the effect of field strength and image contrast. In these configurations, cross-validation was limited to subjects within the corresponding cohort, with all subjects from the other cohort reassigned to the test set of the first fold to assess out-of-domain performance. Next, we trained a model using the back-transformed native-space labels c-FR_M to determine how effectively the network can learn directly in the unrotated canonical FR coordinate system without any cine-based reorientation. Finally, we devised a data-augmentation strategy by extending our 3D-propagation approach into a 4D-propagation method, propagating the semi-automatic 3D FR_M segmentations from ED and ES to all intermediate cardiac phases, creating an all-cardiac phases (ap) segmentation, which is hereafter denoted ap-FR_M. Although this volume-to-volume propagation uses the same combination of diffeomorphic registration formalism, its application across whole 3D volumes, rather than across 2D slices, has never, and practically cannot, be validated against full 4D manual ground truth, since exhaustive annotation of every phase is prohibitive. Accordingly, 4D-propagation served solely as realistic data augmentation, drawing on prior work that uses non-linear registration to enrich volumetric training data [32] and expanding our training set by roughly eightfold ($\approx$16–20 volumes per subject instead of ED and ES only) without additional manual effort.

Upon inference completion, for all models, a post-processing step retained only the largest connected component per label, eliminating spurious islands of misclassification. Automatic segmentations produced by the models trained on FR_M are hereafter referred to as FR_A; those from D_1-FR_M as D_1-FR_A; from D_2-FR_M as D_2-FR_A; from c-FR_M as c-FR_A; and from ap-FR_M as ap-FR_A.

2.5. Validation metrics

Validation of our 4D CMR processing framework relied on three complementary categories of metrics—geometric, clinical, and consistency—to fully characterize performance in both spatial and functional domains. Although the DL model automatically generates segmentations for every cardiac phase, all reported metrics were calculated only at ED and ES, since manual ground truth segmentation exists solely for those two phases.

To assess geometric accuracy, we compared paired segmentations defined in the same spatial frame (e.g., FR_M vs. FR_A, c-FR_M vs. c-FR_A,

etc.) using established medical image segmentation metrics [33]. Volumetric overlap was quantified by the Dice similarity coefficient (DSC), which measures the degree of spatial overlap (intersection) between automatic and ground-truth segmentation masks. Boundary agreement was characterized based on the minimum distances between each point on the automatic and ground-truth boundaries and their closest counterparts. The maximum of these minimum distances defines the Hausdorff distance (HD), while their average in both directions defines the average symmetric surface distance (ASSD). We further evaluated systematic bias in volume estimation via the relative volume difference (RVD) defined as $RVD = |V_A - V_M|/V_M$, where V_A and V_M are the automatic and reference volumes, respectively. All geometric metrics were computed over the pooled set of ED and ES volumes rather than averaging phase-specific results, ensuring each voxel and each boundary discrepancy contributes equally to the overall assessment.

To evaluate clinical fidelity, we compared end-diastolic volume (EDV), end-systolic volume (ESV), and ejection fraction (EF), defined as $EF = (EDV - ESV)/EDV \times 100\%$, for both the left and right ventricular blood pools. Because these parameters are inherently independent of imaging plane selection and voxel grid (being volume integrals of the segmented labels), they could be computed and compared across any pair of segmentation sets (e.g., Cine_M vs. FR_M vs. FR_A), providing a direct measure of cardiac anatomical and functional agreement.

To ensure physiological plausibility within each segmentation, we computed two consistency metrics. The left-ventricular myocardial volume mismatch ε_{LVM} was computed as the absolute difference between the ED and ES LVM volumes (EDV_LVM and ESV_LVM, respectively) divided by the average of these two volumes:

$$\varepsilon_{LVM} = \frac{|EDV_{LVM} - ESV_{LVM}|}{(EDV_{LVM} + ESV_{LVM})/2} \quad (1)$$

This magnitude-based formulation serves as an internal consistency indicator for ED–ES LVM volume variability. In addition, stroke volumes (SV = EDV − ESV) of the left (LVSV) and right (RVSV) ventricles

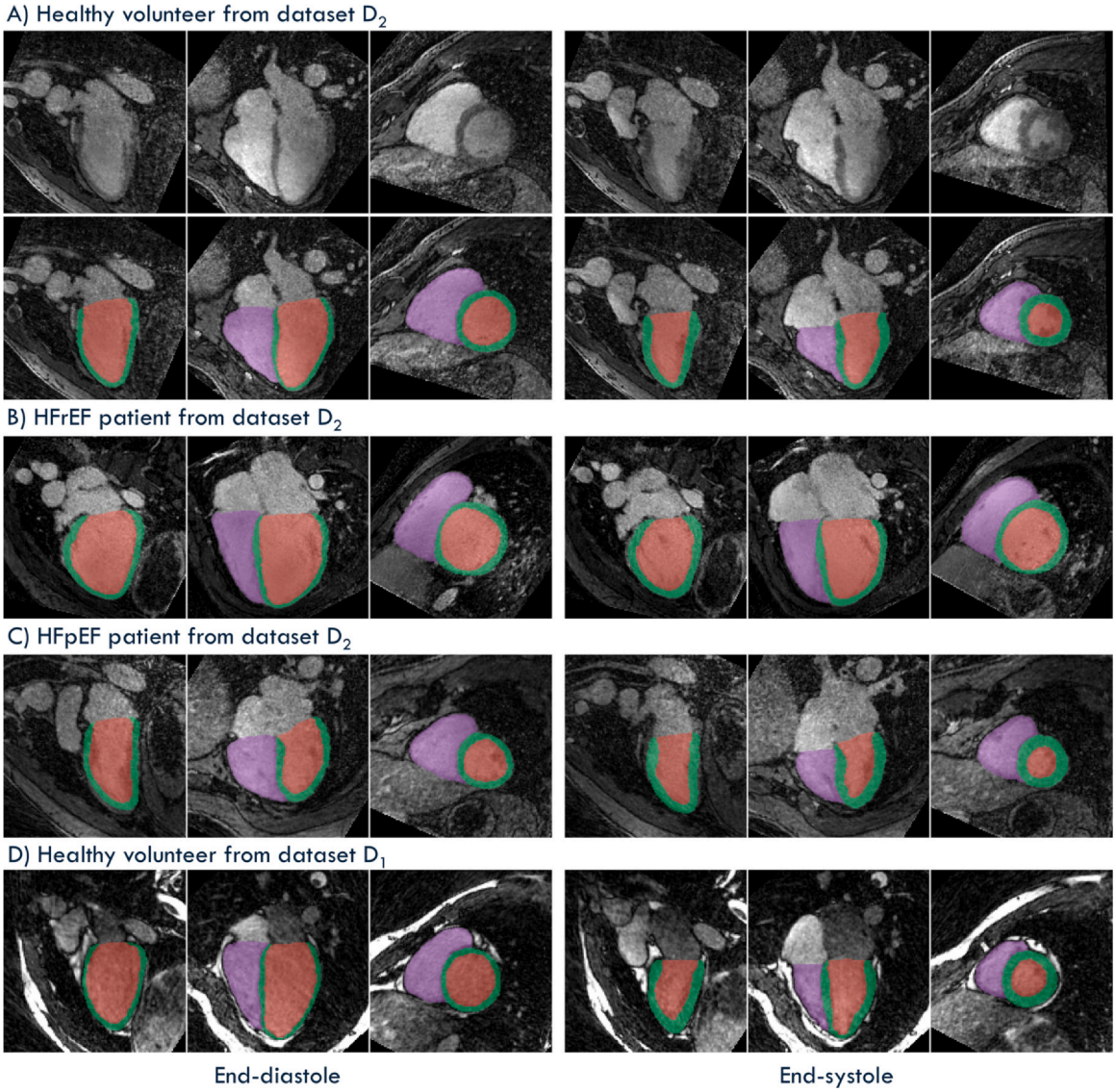

Fig. 2. DL-based automatic segmentation (FR_A) of spatially isotropic 4D FR images at end-diastole (left) and end-systole (right). Examples are shown for A) a healthy volunteer in their 60 s, B) a patient in their 30 s with HFrEF and C) a patient in their 60 s with HFpEF, all scanned at 3T (dataset D_2), as well as D) a healthy volunteer in their 20 s scanned at 1.5T (dataset D_1). The FR images are displayed in pseudo two-chamber, pseudo four-chamber, and short-axis views, with the three orthogonal planes corresponding to those used in the semi-automatic segmentation that served as ground truth for DL-based training. Segmented structures include the right ventricle blood pool (purple), left ventricle myocardium (green), and left-ventricular blood pool (red). In panel D (1.5T bSSFP), the blood in the left-sided chambers appears darker due to a combination of inflow effects and slab selective excitation. *DL* deep learning, *FR* free-running, *4D* four-dimensional, *HFrEF* heart failure with reduced ejection fraction, *HFpEF* heart failure with preserved ejection fraction, *bSSFP* balanced steady-state free precession

were compared in healthy subjects only; agreement between LVSV and RVSV provided a second metric of intrinsic physiological validity.

All results are reported as mean ± standard deviation. Measurement agreement was evaluated by Bland–Altman analysis, providing bias and limits of agreement (LOA). Reproducibility and precision were quantified using the intraclass correlation coefficient with a (2, 1) formula [34] and the corresponding standard error of measurement (SEM). Two-sided paired Student's t-tests were used to compare normally distributed paired measurements (normality assessed with Shapiro–Wilk test). For ε_{LVM}, a one-way repeated-measures analysis of variance (ANOVA) was performed to compare the different segmentation approaches, with post hoc pairwise comparisons carried out using Tukey's HSD procedure (with familywise $\alpha = 0.05$).

3. Results

For ground truth generation, our 3D-propagation scheme reduced the need for manual contouring to 28 ± 5% of ED and ES slices. Specifically, at ED, 17 ± 5 RV and 18 ± 3 LV slices were manually segmented out of a total of 64 ± 7 RV and 68 ± 7 LV slices. At ES, 16 ± 4 RV and 17 ± 3 LV slices were contoured from total volumes of 50 ± 8 RV and 58 ± 8 LV slices. After propagation, fewer than 5% of slices required minor correction, adding, removing, or adjusting labels, to enforce myocardial volume consistency within the expected physiological tolerance and stroke-volume agreement.

All DL trainings converged in 80 epochs (~220 s per epoch, total ~5 h), and all DL models consistently labeled cardiac structures across every cardiac phase (Supplementary Figure M1). Inference on a full 4D series required ~1 min per subject. Only ED and ES frames were retained for subsequent analysis (Fig. 2). FR images from all 35 subjects were successfully segmented. Larger local segmentation failures occurred in 2 cases with atypical anatomy or implanted devices, where basal slices or RV trabeculae were sometimes mis-segmented (Supplementary Figs. S1–S2).

3.1. Geometric metrics

FR_A closely matched FR_M across all geometric metrics (Fig. 3). The DSC between FR_A and FR_M was 0.94 ± 0.01 for LVB, 0.86 ± 0.02 for LVM, and 0.92 ± 0.01 for RVB. Boundary errors remained on the order of a single voxel for ASSD at 1.0 ± 0.1 mm (LVB), 0.9 ± 0.1 mm (LVM), and 1.2 ± 0.3 mm (RVB), while HD values were 6.2 ± 1.3 mm, 7.1 ± 2.0 mm, and 13.8 ± 6.5 mm, respectively. Manual-automatic volume bias was minimal, with RVD of 2.7 ± 2.1% for LVB, 5.8 ± 3.7% for LVM, and 4.5 ± 2.8% for RVB. The subject responsible for the HD outlier for RV segmentation is shown in Fig. S1.

3.1.1. Dataset-specific generalization

Models trained on all data (FR_A) and on individual datasets (D_1-FR_A, D_2-FR_A) achieved similar high overlap and low volume bias when evaluated within their training cohort (Table 3). In contrast, cross-cohort performance dropped sharply for both overlap and bias: D_1-FR_A on D_2-FR_M yielded DSC < 0.65 and RVD > 32.0%, while D_2-FR_A on D_1-FR_M resulted in DSC < 0.63 and RVD > 22.2% for all regions of interest.

3.1.2. Orientation and data augmentation effects

When applied to the native-space reference (c-FR_M, the canonical orientation in which the data was acquired), the FR_A model, trained only on SAX-reoriented images, showed a slight decrease in both DSC and RVD (Table 4), reflecting reduced ability of the models to generalize to unseen orientations. However, training and evaluating directly in native canonical space (c-FR_A on c-FR_M) restored performance to levels equivalent to those achieved on reoriented volumes (Table 4).

The all-cardiac phases model (ap-FR_A) evaluated on FR_M resulted in high DSC and RVD that were very similar to FR_A on FR_M (Table 4), demonstrating that this data augmentation strategy has a negligible effect on geometric accuracy.

3.2. Clinical metrics

Functional agreement between FR_M and FR_A was very high for the left ventricle (Fig. 4). Bland–Altman analysis of EDV revealed very high reproducibility (ICC = 0.99, SEM = 3.7 mL), though the small bias was statistically significant (p = 0.04), the LOA remained tight (−8.3; 12.0 mL). End-systolic volumes agreed even more closely (ICC = 1.00, SEM = 3.5 mL), and ejection fraction showed high concordance (ICC = 0.98, SEM = 2.0%, p = 0.34, LOA = −5.0; 5.9%). Right ventricular function followed a similar pattern, with high reproducibility (EDV ICC = 0.99, SEM = 6.4 mL; ESV ICC = 0.97, SEM = 7.4 mL; EF ICC = 0.93, SEM = 3.3%) and negligible bias, but wider LOA.

In comparison, agreement between FR_M and $Cine_M$ was high for the left ventricle, with ICCs of 0.96, 0.98, and 0.91 for EDV, ESV, and EF, respectively, but exhibited wider LOA. For the right ventricle, the LOA were even wider, and ICCs were lower (EDV ICC = 0.88, ESV ICC = 0.92, EF ICC = 0.66), reflecting reduced measurement reproducibility.

The clinical metrics for the c-FR_A and ap-FR_A configurations show that c-FR_A introduces only minimal degradation in reliability and

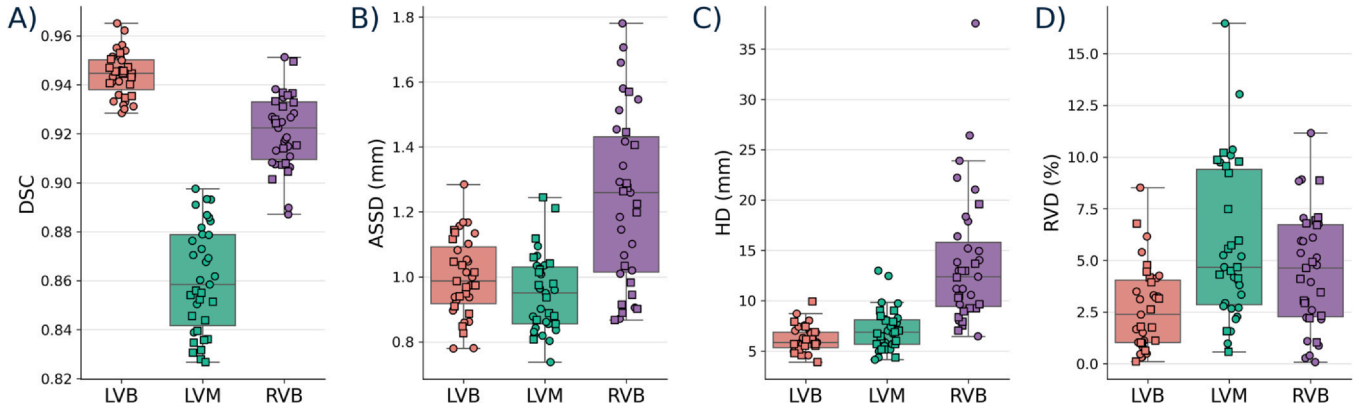

Fig. 3. Geometric comparison between the ground truth segmentation (FR_M) and the DL-based automatic segmentation (FR_A) of the 4D FR images using metrics calculated from the combined end-diastolic and end-systolic frames. A) DSC, B) ASSD, C) HD, and D) RVD for the LVB, LVM, and RVB. Square markers represent subjects from dataset D_1, and circle markers indicate subjects from dataset D_2. Boxplots use the Tukey definition (Q1–Q3 box, median line, whiskers to 1.5 × IQR). All data points, including outliers, were included in the statistics. *FR* free-running, *DL* deep learning, *4D* four-dimensional, *DSC* dice similarity coefficients, *ASSD* average symmetric surface distance, *HD* Hausdorff distance, *RVD* relative volume difference, *LVB* left-ventricular blood pool, *LVM* left-ventricular myocardium, *RVB* right ventricular blood pool

Table 3
Geometric accuracy and volume bias (DSC and RVD) for models trained on different datasets (D_1, D_2, and combined) and evaluated on FR_M segmentations from each cohort

Trained on:		FR_A		D_1-FR_A		D_2-FR_A	
Validated against:		D_1-FR_M	D_2-FR_M	D_1-FR_M	D_2-FR_M	D_1-FR_M	D_2-FR_M
DSC	LVB	0.94 ± 0.00	0.94 ± 0.01	0.94 ± 0.01	0.62 ± 0.29	0.67 ± 0.14	0.94 ± 0.01
(–)	LVM	0.84 ± 0.01	0.87 ± 0.02	0.83 ± 0.01	0.49 ± 0.22	0.43 ± 0.16	0.87 ± 0.02
	RVB	0.92 ± 0.01	0.92 ± 0.02	0.92 ± 0.02	0.59 ± 0.20	0.24 ± 0.30	0.92 ± 0.02
RVD	LVB	2.6 ± 1.9	2.7 ± 2.1	2.3 ± 1.9	29.8 ± 25.5	36.3 ± 18.3	3.1 ± 2.1
(%)	LVM	6.2 ± 3.0	5.4 ± 4.1	6.0 ± 3.6	33.5 ± 43.7	22.1 ± 21.9	6.3 ± 4.6
	RVB	4.5 ± 2.2	4.5 ± 3.1	4.8 ± 3.2	48.3 ± 47.3	66.9 ± 23.0	5.4 ± 3.6

FR_A denotes the model trained on the combined D_1 and D_2 cohort; D_1-FR_A and D_2-FR_A refer to models trained exclusively on D_1 or D_2, respectively. The combined model generalizes well across datasets, while single-dataset models perform best on their own cohort but poorly on the other. Data are reported as means ± standard deviation.

Table 4
Geometric accuracy and volume bias (DSC and RVD) for different training configurations

Trained on:		FR_A	FR_A	c-FR_A	ap-FR_A
Validated against:		FR_M	c-FR_M	c-FR_M	FR_M
DSC	LVB	0.94 ± 0.01	0.92 ± 0.02	0.94 ± 0.01	0.94 ± 0.01
(–)	LVM	0.86 ± 0.02	0.82 ± 0.03	0.85 ± 0.02	0.85 ± 0.02
	RVB	0.92 ± 0.01	0.86 ± 0.04	0.91 ± 0.02	0.92 ± 0.02
RVD	LVB	2.7 ± 2.0	6.8 ± 4.1	3.7 ± 2.8	4.0 ± 3.2
(%)	LVM	5.8 ± 3.7	8.4 ± 7.1	7.3 ± 4.5	7.1 ± 4.0
	RVB	4.5 ± 2.8	11.4 ± 10.4	5.0 ± 4.3	4.7 ± 3.1

Models were trained on native-space data (c-FR_A), reoriented FR data (FR_A), or FR_M with 4D propagation-based augmentation (ap-FR_A), and validated against corresponding ground truth segmentations (c-FR_M or FR_M). FR_A applied to native-space data showed a slight performance drop, while training directly in native space restored performance. The all-cardiac phases model ap-FR_A achieved comparable performance to FR_A, indicating that the data augmentation strategy has a negligible effect on geometric accuracy. Data are reported as means ± standard deviation.

accuracy compared to FR_A (with all ICC > 0.90), whereas ap-FR_A exhibits a more pronounced, but still modest, increase (with all ICC > 0.93) in LOA and bias (Supplementary Fig. S3). For both DL configurations, the lowest reproducibility was observed for RV EF, although the bias remained below 1.5% with LOA around 10%.

3.3. Physiological consistency metrics

The myocardial volume mismatch between systole and diastole ε_{LVM} (Fig. 5) was the highest for $Cine_M$ (7.1 ± 4.5%), followed by FR_M (5.6 ± 3.5%), with no significant difference between these two methods ($p = 0.28$). ε_{LVM} decreased non-significantly when moving from FR_M to FR_A (4.0 ± 2.4%; $p = 0.24$). c-FR_A (4.4 ± 3.4%) also did not differ from FR_A ($p = 0.98$). The lowest volume mismatch was observed for ap-FR_A (2.6 ± 1.9%), which was smaller than all other methods. Among automatic methods, it was the only one that differed significantly from FR_M ($p = 0.002$).

LV–RV stroke volume agreement remained high but exhibited progressively greater variability from $Cine_M$ to FR_M to FR_A (Fig. 6). $Cine_M$ yielded an ICC of 0.94 with a modest bias of 2.8 mL ($p = 0.04$). Agreement for FR_M was similarly high (ICC = 0.91) but with increased bias (6.4 mL). FR_A showed wider LOA, a lower ICC of 0.83, and a bias of 7.1 mL, indicating that although all methods preserve stroke-volume parity, automatic segmentation introduces slightly greater variability. LV–RV stroke volume agreement for c-FR_A and ap-FR_A yielded ICCs, LOA, and biases nearly identical to FR_A (Supplementary Fig. S4).

4. Discussion

In this study, we introduced and validated a deep-learning-based segmentation approach specifically tailored to isotropic dynamic (3D + t) FR cardiac MRI. Unlike existing segmentation methods designed for static 3D [13,15,16] or multi-slice 2D [11,12,14] acquisitions, this approach fully leverages the unique properties of FR imaging, including high isotropic spatial resolution, extensive spatial coverage, and comprehensive temporal dynamics, while eliminating the need to acquire predefined cine planes.

Our main method FR_A demonstrated excellent geometric accuracy across evaluated cardiac structures, achieving high DSC scores of 0.94 ± 0.01 (LVB), 0.86 ± 0.02 (LVM), 0.92 ± 0.02 (RVB). The corresponding HD values (6.2 ± 1.3 mm, 7.4 ± 2.6 mm, and 13.8 ± 6.5 mm) were in line with those reported in similar segmentation tasks [11]. Limited volume bias was observed with RVD values of 2.7 ± 2.0% (LVB), 5.7 ± 3.7% (LVM), and 4.5 ± 2.8% (RVB). Direct comparisons with prior studies are inherently complex due to the unique isotropic and dynamic nature of 4D FR datasets. Nonetheless, our achieved geometric metrics meet or exceed the current state-of-the-art reported for widely used CMR datasets. Bai et al. reported DSC scores of 0.94 for LVB, 0.88 for LVM, and 0.90 for RVB on the UK Biobank dataset [12], similar to our results, despite their images having lower through-plane resolution and inter-slice gaps. Using the ACDC challenge data, DSC values of 0.95 (LVB), 0.91 (LVM), and 0.92 (RVB) for top-performing methods were reported [11,14], though again with non-isotropic spatial resolution. Similarly, Bustamante et al. reported DSC values of 0.91 for LV and 0.89 for RV using 4D flow data [35], albeit with lower anatomical clarity and no myocardial segmentation due to limited contrast. On the MMWHS challenge, which includes static 3D whole-heart MR data, reported performances are generally lower for RVB and LVM: Muffoletto et al. achieved DSC scores of 0.87 (LVB), 0.69 (LVM), and 0.74 (RVB) using semi-supervised domain adaptation methods [15], while Wang et al. reported values of 0.86 (LVB), 0.74 (LVM), and 0.85 (RVB) with a two-stage U-Net architecture [16]. Notably, our method also exceeds the DSC values reported for intra-expert variability on the manual MMWHS annotations [13], although they remain slightly lower than the inter-expert variability reported in the ACDC challenge [11]. Thus, despite the higher dimensional complexity, our method achieves comparable or higher accuracy, affirming its potential clinical relevance.

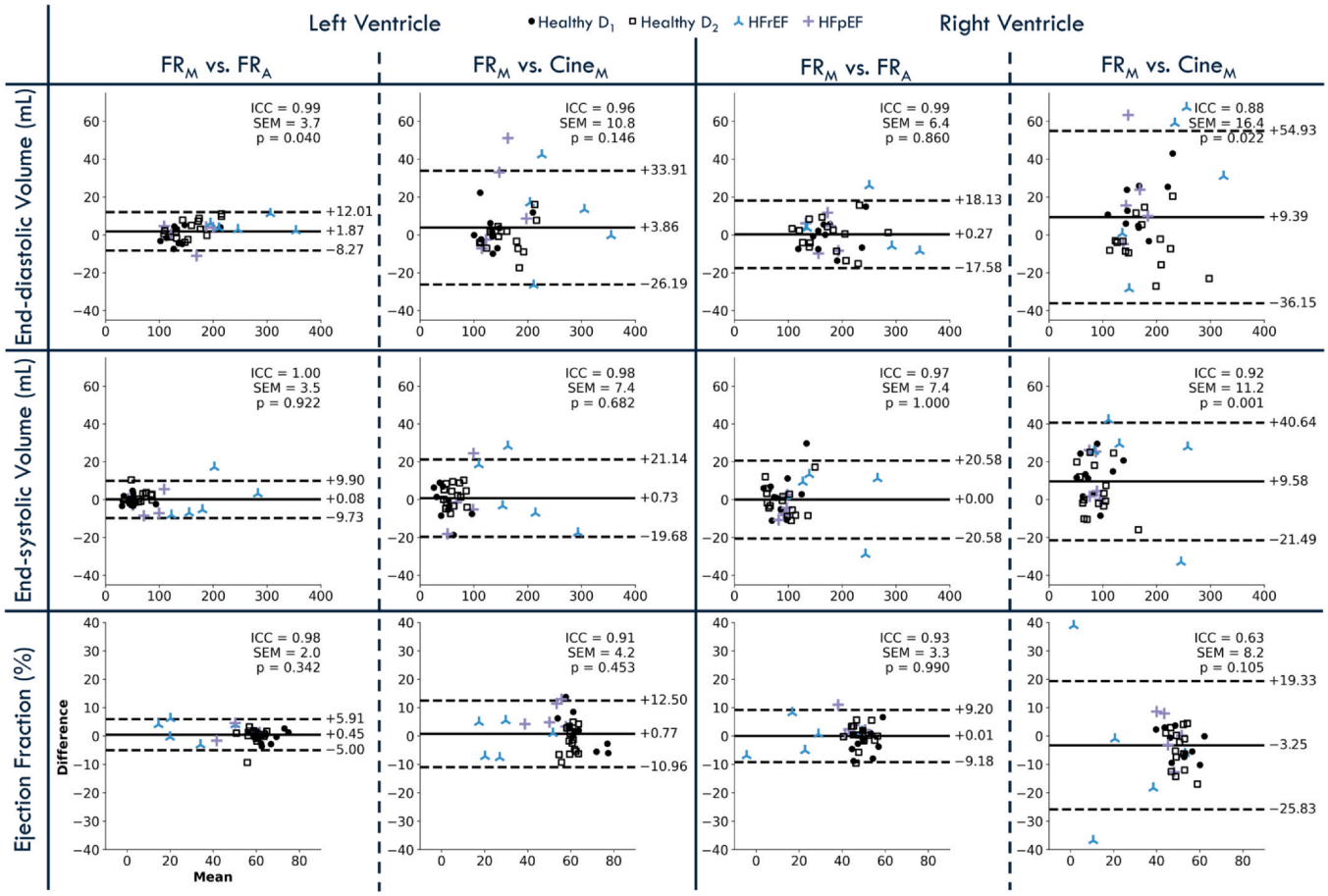

Fig. 4. Clinical metrics comparison through volumetric and functional agreement. Bland–Altman analysis comparing semi-automatic segmentation (FR_M) and deep learning-based automatic segmentation (FR_A) of FR images, as well as FR_M and manual segmentation of cine images ($Cine_M$), for end-diastolic volume, end-systolic volume, and ejection fraction in both left and right ventricles. The x-axis represents the average of the two methods being compared, while the y-axis indicates the difference between them. The central solid line shows the bias, and dashed lines indicate the limits of agreement (bias ± 1.96 standard deviations). ICC, SEM, and p-values for each comparison are also reported. Healthy volunteers are represented by black dots (square markers for dataset D_1, circle markers for dataset D_2), patients with HFrEF by blue crosses, and those with HFpEF by purple crosses. *ICC* Intraclass correlation coefficients, *SEM* standard error of measurement, *HFrEF* heart failure with reduced ejection fraction, *HFpEF* heart failure with preserved ejection fraction

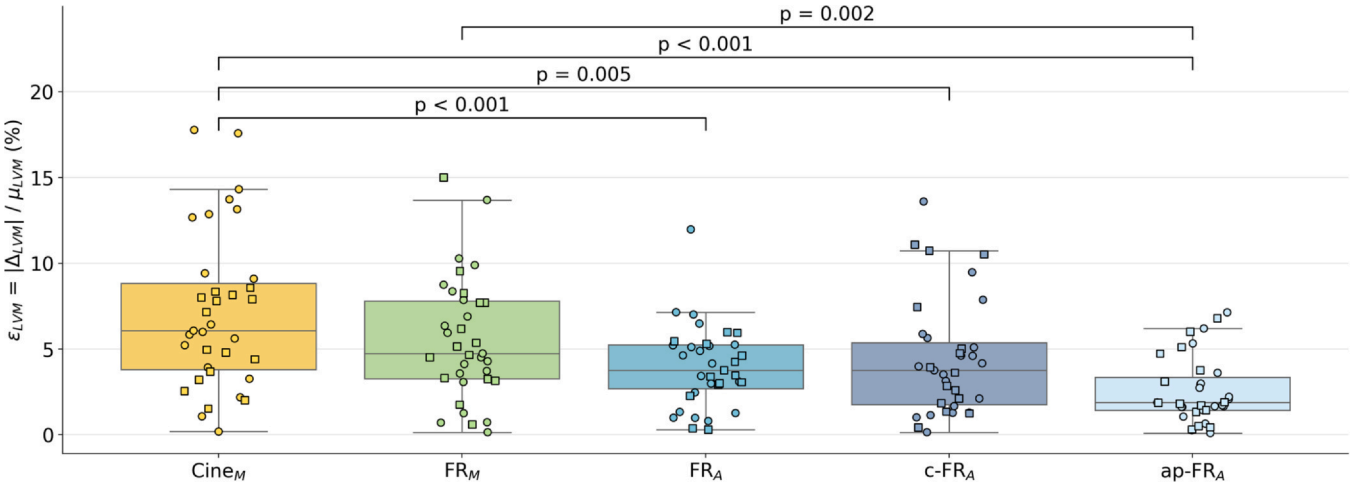

Fig. 5. Systolic–diastolic myocardial volume mismatch. LVM volume mismatch ε_{LVM}, defined as the absolute difference between end-diastolic and end-systolic volumes (Δ_{LVM}) divided by their mean (μ_{LVM}), is shown for manual segmentation of cine images ($Cine_M$), semi-automatic segmentation of 4D FR images (FR_M), and three DL training dataset configurations (FR_A, c-FR_A, and ap-FR_A). No significant difference was observed between $Cine_M$ and FR_M ($p = 0.28$), while all FR-based models yielded significantly smaller errors than $Cine_M$. Among automatic methods, ap-FR_A yielded the lowest ε_{LVM} and was the only one significantly different from FR_M ($p = 0.002$). Square markers represent subjects from dataset D_1, and circle markers indicate subjects from dataset D_2. *LVM* left-ventricular myocardium, *FR* free-running

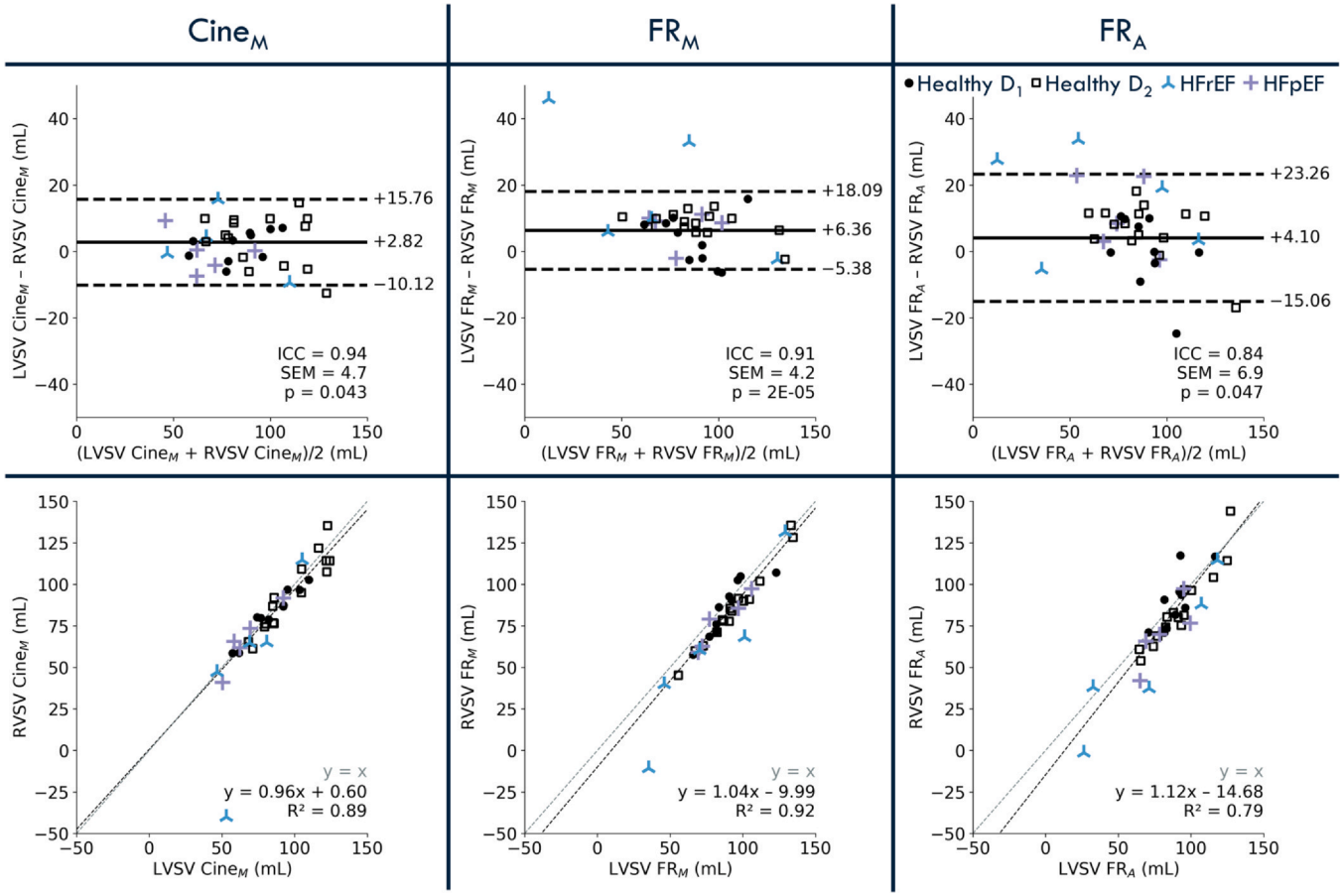

Fig. 6. Stroke volume agreement. Bland–Altman analyses (top) and correlation plots (bottom) comparing RVSV and LVSV, computed from manual segmentation of cine images ($Cine_M$) and from semi-automatic (FR_M) and deep learning-based automatic segmentation (FR_A) of 4D FR images. In the Bland–Altman analysis, the central horizontal line represents the bias, and dashed lines indicate the limits of agreement (bias ± 1.96 SD). Agreement levels were calculated using only healthy volunteer data, since cardiac regurgitation was unknown for the patients. ICC, SEM, and p-values for each comparison are also reported. Healthy volunteers are represented by black dots (square markers for dataset D_1, circle markers for dataset D_2), patients with HFrEF by blue crosses, and those with HFpEF by purple crosses. *RVSV* stroke volumes of the right ventricle, *LVSV* stroke volumes of the left ventricle, *4D* four-dimensional, *ICC* intraclass correlation coefficients, *SEM* standard error of measurement, *HFrEF* heart failure with reduced ejection fraction, *HFpEF* heart failure with preserved ejection fraction

We extended validation beyond geometric overlap by directly assessing key clinical metrics. Comparing FR_A to FR_M for the LV, FR_A achieved an ICC > 0.99 for both EDV and ESV, with biases of 1.9 mL (LOA: −8.3; 12.0 mL) and 0.1 mL (−9.7; 9.9 mL), respectively. EF also showed excellent agreement (ICC = 0.98, bias = 0.45%, LOA: −5.0; 5.9%). These results closely mirror established intra-observer variability for 2D cine LV segmentation, where prior work [36] reports an EDV bias of 1.9 mL (LOA: −7.9; 11.7 mL), an ESV bias of 0.6 mL (LOA: −6.7; 8.0 mL), and an EF bias of 0.2% (LOA: −4.9; 4.6%). For the RV, the clinical robustness of FR_A remained high (ICC: EDV = 0.98, ESV = 0.97, EF = 0.93) with minimal biases (0.27 mL, 0.00 mL, 0.01%) but wider LOA (EDV: −17.6; 18.1 mL; ESV: −20.6; 20.6 mL; EF: −9.2; 9.2%), reflecting the known challenges of RV delineation.

Consistent with previous studies [10,36], segmentation accuracy was notably higher for the LV than the RV. This discrepancy likely results from the manual segmentation process that is primarily performed in the short-axis orientation. This orientation is optimized for LV delineation and complicates precise RV segmentation, particularly at the basal RV region. The complexity of the manual RV segmentation impacts both the training and validation of the automatic method. Future RV segmentations may need to be performed in a dedicated oblique orientation.

We further introduced two physiological consistency metrics. We demonstrated lower myocardial volume mismatch between the ED and ES phases (ε_{LVM}) in automatic segmentations compared to manual methods. Conversely, LV–RV stroke volume agreement was slightly reduced in the automatic approach, reflecting the inherent challenges of RV segmentation. Higher agreement in manual segmentations can be explained by the fact that clinicians frequently use this type of physiological cue to iteratively refine contours whenever appropriate [24,28]. While this strategy improves consistency, it also substantially increases analysis time. In contrast, our automated method provides accurate results within approximately one minute per dataset. While these physiological consistency metrics are not directly used in clinical decision making, they are valuable indicators of physiological plausibility and could support future automated quality control.

When trained on the combined D_1 and D_2 cohort, the FR_A model achieved high geometric accuracy compared to both D_1-FR_M and D_2-FR_M, demonstrating effective learning across different contrasts and field strengths. In contrast, models trained exclusively on a single dataset (D_1-FR_A or D_2-FR_A) performed well on their own cohort but showed significant accuracy drops when applied to the other, revealing limited cross-cohort transfer of the training applicability. Importantly, within each dataset, a model trained on that dataset alone matched the performance of the combined-training model on that same data. These results highlight that broad generalization requires diverse, multi-center, multi-field-strength, and multi-contrast FR datasets, as already explored by several recent studies [8,9,37–39], to ensure robust performance across varied clinical scenarios, whereas single-cohort models cannot reliably segment out-of-domain FR images.

Training the segmentation model directly in the native orientation (i.e., the canonical orientation in which each FR dataset was acquired) with c-FR_A effectively addressed the rotation sensitivity observed when applying the FR_A model to native-space volume (c-FR). Because c-FR_A learns in each subject's intrinsic cardiac orientation rather than in the shared SAX template used for FR_A, it confirmed that segmentation accuracy does not depend on cine-derived orientations. Specifically, c-FR_A on c-FR_M achieved DSC scores of 0.94 ± 0.01 (LVB), 0.85 ± 0.02 (LVM), and 0.91 ± 0.02 (RVB), with RVD values of 3.7 ± 2.8%, 7.3 ± 4.5%, and 5.0 ± 4.3%, closely matching FR_A performance on rotated FR_M data. Agreement on clinical and physiological consistency metrics was also maintained. These findings show that native-space training provides equivalent accuracy while removing the need for reorientation, simplifying future workflows by eliminating dependence on cine-prescribed oblique orientations, which can instead be generated post hoc during analysis.

Augmenting the training dataset with 4D propagation-based data augmentation (ap-FR_A) had minimal impact on per-phase accuracy and did not affect clinical accuracy. This evaluation is limited to ED and ES frames; ap-FR_A may improve segmentation accuracy on intermediate cardiac phases, though this remains unverified without all-cardiac phase ground truth segmentation. However, ap-FR_A showed improved internal temporal consistency, reducing the LVM volume mismatch ε_{LVM} from 4.0 ± 2.4% (FR_A) to 2.6 ± 1.9%. As the true physiological myocardial volume change remains debated and is likely small, ε_{LVM} reductions should be interpreted as methodological improvements rather than physiological claims. This reduction likely reflects the broader distribution of cardiac shapes seen during training, as intermediate phases expose the network to a wider range of myocardial geometries. Such improvement underscores the potential value of 4D propagation in scenarios where temporal coherence across phases is particularly important, such as myocardial strain analysis, while remaining optional for standard volumetric or functional assessments.

An important contribution of this study is the semi-automatic generation of ground truth segmentations, bridging the gap between fully manual annotations and the extensive datasets required for DL training. Although not fully manual, this semi-automatic approach relies on a validated method, making comprehensive annotation feasible for this study. The extent of segmentation required for accurate validation would have been impractical to achieve entirely manually. We publicly shared the code for our 3D and 4D propagation tools and FR reorientation script, alongside example datasets, via an open-access GitLab repository (*gitlab.com/augustin-c-ogier/segpropa*).

Although vision transformers [40] and foundation models are gaining attention in medical imaging, we selected nnU-Net for its proven performance with relatively small datasets and its practicality. It consistently outperformed other methods in several challenges [30] and adapts well to varied medical datasets without manual tuning. In contrast, transformer-based models require large-scale pretraining and remain largely untested on high-resolution 3D+t medical images. nnU-Net allowed efficient training from scratch on our FR data, making it a robust and accessible solution. Future work may explore transformers as annotated FR datasets grow.

Several limitations should nonetheless be recognized. Firstly, our segmentation included only LVB, LVM, and RVB. Comprehensive cardiac assessments would benefit from incorporating additional structures, such as the atrial chambers and valve structures, which would require dedicated additional annotations but could further enhance boundary definition and broaden clinical applicability. Secondly, while our model can be used to segment all cardiac phases, validation was limited to ED and ES frames and extending the cardiac assessment to motion-related abnormalities or valve dysfunction is currently constrained by the absence of full-cycle 4D ground truth annotations. A potential solution lies in further validating our 4D propagation method, which could serve as a reliable temporal reference and support the integration of advanced temporal architectures such as long short-term memory [41]. However, training on full 4D volumes would pose substantial memory and computational challenges. Likewise, our analysis focused solely on the end-expiratory phase of the reconstructed 5D data, and validation across the full 5D space, encompassing both cardiac and respiratory dynamics, remains methodologically complex, but a promising direction for future research.

Additionally, while tested across two magnetic field strengths and imaging contrasts, future work should assess performance across broader imaging conditions, including different contrast-enhanced protocols [37,38], low-field systems [9], more varied patients cohorts, and other acquisition variants using modified pulse sequences [8,39]. Although the model performed well across a range of patient cases, its training dataset included only a limited number of pathological subjects, and reduced accuracy was observed in individuals with markedly atypical anatomy. Future work will require larger and more diverse pathological cohorts, such as patients with congenital heart disease [38], to fully assess robustness in populations with substantial anatomical variability.

In conclusion, free-running imaging offers substantial advantages over conventional 2D cine MRI, including improved ease-of-use and time efficiency, isotropic spatial resolution, and rich spatiotemporal information. However, the analytical complexity of processing 5D data has long posed a barrier to clinical adoption. By introducing an accurate, rapid, and scalable segmentation framework, we help mitigate this barrier and enable reliable extraction of anatomical and functional metrics from full 3D + t FR data. Notably, our validation extends beyond standard geometric metrics to include clinically meaningful indices and physiological consistency measures, offering a more complete evaluation than most segmentation studies to date. This comprehensive assessment supports the reliability of our approach not only for anatomical delineation but also for functional analysis. By enabling efficient analysis of dynamic whole-heart data, this work may contribute to the integration of FR CMR into clinical workflows with limited cardiac MR expertise.

Funding

This study was funded by the Swiss National Science Foundation under grant CRSII5_202276 (to RBvH, RH, JR, and Philippe Meyer), as well as in part through SNSF grants 320030B_201292, 320030_173129, and 326030_150828 (to MS).

Author contributions

Augustin C. Ogier: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Salomé Baup:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Data curation. **Gorun Ilanjian:** Investigation. **Aisha Touray:** Investigation. **Angela Rocca:** Writing – review & editing, Resources, Investigation, Data curation. **Jaume Banús:** Writing – review & editing, Software, Methodology. **Isabel Montón Quesada:** Writing – review & editing, Visualization, Methodology. **Martin Nicoletti:** Writing – review & editing, Methodology. **Jean-Baptiste Ledoux:** Writing – review & editing, Resources, Investigation. **Jonas Richiardi:** Writing – review & editing, Funding acquisition, Formal analysis. **Robert J. Holtackers:** Writing – review & editing, Resources, Investigation. **Jérôme Yerly:** Writing – review & editing, Software, Methodology. **Matthias Stuber:** Writing – review & editing, Funding acquisition. **Roger Hullin:** Writing – review & editing, Resources, Funding acquisition. **David Rotzinger:** Writing – review & editing, Validation, Supervision, Investigation. **Ruud B. van Heeswijk:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Ethics approval and consent

Ethics approval was obtained from the Ethics Committee of the Canton of Vaud (CER-VD) of Switzerland under reference numbers 2022–00934 and 2021–02458. All participants provided written informed consent to participate prior to enrollment.

Data and code availability

The code used for both the 3D and 4D propagation methods, as well as the reorientation of free-running images to match 2D cine views, is available on GitLab at *gitlab.com/augustin-c-ogier/segpropa*, along with a representative 3T gadolinium-based contrast agent (GBCA)–enhanced dataset for testing and demonstration purposes. The nnU-Net framework is available from the original authors at *github.com/MIC-DKFZ/nnUNet*. The majority of datasets cannot be shared due to restrictions imposed by local ethics regulations. Consequently, the trained deep learning models, which rely on this data, are also not publicly available. However, both can be made available upon reasonable request and pending appropriate approvals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jocmr.2025.102677.

References

[1] Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1736–88. https://doi.org/10.1016/S0140-6736(18)32203-7.

[2] Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. J Cardiovasc Magn Reson 2020;22(1):17. https://doi.org/10.1186/s12968-020-00607-1.

[3] Yang K, Cui C, Teng F, Yin G, An J, Yang X, et al. Full free-breathing cardiovascular magnetic resonance imaging: enhancing efficiency and image quality in clinical practice. J Cardiovasc Magn Reson 2025;27(2):101955. https://doi.org/10.1016/j.jocmr.2025.101955.

[4] Küstner T, Fuin N, Hammernik K, Bustin A, Qi H, Hajhosseiny R, et al. CINENet: deep learning-based 3D cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions. Sci Rep 2020;10(1):13710. https://doi.org/10.1038/s41598-020-70551-8.

[5] Christodoulou AG, Shaw JL, Nguyen C, Yang Q, Xie Y, Wang N, et al. Magnetic resonance multitasking for motion-resolved quantitative cardiovascular imaging. Nat Biomed Eng 2018;2(4):215–26. https://doi.org/10.1038/s41551-018-0217-y.

[6] Di Sopra L, Piccini D, Coppo S, Stuber M, Yerly J. An automated approach to fully self-gated free-running cardiac and respiratory motion-resolved 5D whole-heart MRI. Magn Reson Med 2019;82(6):2118–32. https://doi.org/10.1002/mrm.27898.

[7] Holtackers RJ, Ogier AC, Romanin L, Tenisch E, Montón Quesada I, Van Heeswijk RB, et al. How low can we go? The effect of acquisition duration on cardiac volume and function measurements in free-running cardiac and respiratory motion-resolved five-dimensional whole-heart cine magnetic resonance imaging at 1.5T. J Cardiovasc Magn Reson 2025;27(1):101863. https://doi.org/10.1016/j.jocmr.2025.101863.

[8] Ogier AC, Montón Quesada I, Sieber X, Calarnou P, Ledoux J, Milani B, et al. Free-running 5D whole-heart MRI for isotropic cardiac function measurements at 3T without contrast agents. Magn Reson Med 2025;93(6):2386–400. https://doi.org/10.1002/mrm.30469.

[9] Sieber X, Binzel K, Varghese J, Liu Y, Yerly J, Roy CW, et al. Measuring biventricular function and left atrial volume in a single five-dimensional whole-heart cardiovascular magnetic resonance scan at 0.55T. J Cardiovasc Magn Reson 2025;27(1):101906. https://doi.org/10.1016/j.jocmr.2025.101906.

[10] Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep learning for cardiac image segmentation: a review. Front Cardiovasc Med 2020;7:25. https://doi.org/10.3389/fcvm.2020.00025.

[11] Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 2018;37(11):2514–25. https://doi.org/10.1109/TMI.2018.2837502.

[12] Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018;20(1):65. https://doi.org/10.1186/s12968-018-0471-x.

[13] Zhuang X, Li L, Payer C, Štern D, Urschler M, Heinrich MP, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med Image Anal 2019;58:101537. https://doi.org/10.1016/j.media.2019.101537.

[14] Isensee F, Jaeger PF, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In: Pop M, Sermesant M, Jodoin PM, Lalande A, Zhuang X, Yang G, Young A, Bernard O, editors. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Vol 10663. Lecture Notes in Computer Science Springer International Publishing; 2018. p. 120–9. https://doi.org/10.1007/978-3-319-75541-0_13.

[15] Muffoletto M, Xu H, Kunze KP, Neji R, Botnar R, Prieto C, et al. Combining generative modelling and semi-supervised domain adaptation for whole heart cardiovascular magnetic resonance angiography segmentation. J Cardiovasc Magn Reson 2023;25(1):80. https://doi.org/10.1186/s12968-023-00981-6.

[16] Wang C, MacGillivray T, Macnaught G, Yang G, Newby D. A Two-Stage U-Net Model for 3D multi-class segmentation on full-resolution cardiac data. In: Pop M, Sermesant M, Zhao J, Li S, McLeod K, Young A, Rhode K, Mansi T, editors. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Vol 11395. Lecture Notes in Computer Science Springer International Publishing; 2019. p. 191–9. https://doi.org/10.1007/978-3-030-12029-0_21.

[17] Sun X, Cheng LH, Plein S, Garg P, Van Der Geest RJ. Deep learning based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI. J Cardiovasc Magn Reson 2024;26(1):100003. https://doi.org/10.1016/j.jocmr.2023.100003.

[18] Ogier A, Sdika M, Foure A, Le Troter A, Bendahan D. Individual muscle segmentation in MR images: a 3D propagation through 2D non-linear registration approaches. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE,; 2017. p. 317–20. https://doi.org/10.1109/EMBC.2017.8036826.

[19] Meyer P, Rocca A, Banus J, Ogier AC, Georgantas C, Calarnou P, Fatima A, Vallée JP, Deux JF, Thomas A, Marquis J, Monney P, Lu H, Ledoux JB, Tillier C, Crowe LA, Abdurashidova T, Richiardi J, Hullin R, Van Heeswijk RB. The HeartMagic prospective observational study protocol – characterizing subtypes of heart failure with preserved ejection fraction. Cardiovasc Med Prepr Post Online April 10, 2025. https://doi.org/10.1101/2025.04.10.25325567.

[20] Piccini D, Littmann A, Nielles-Vallespin S, Zenge MO. Spiral phyllotaxis: The natural way to construct a 3D radial trajectory in MRI. Magn Reson Med 2011;66(4):1049–56. https://doi.org/10.1002/mrm.22898.

[21] Vincenti G, Monney P, Chaptinel J, Rutz T, Coppo S, Zenge MO, Schmidt M, Nadar MS, Piccini D, Chèvre P, Stuber M, Schwitter J. Compressed Sensing Single–Breath-Hold CMR for Fast Quantification of LV Function, Volumes, and Mass. JACC Cardiovasc Imaging 2014;7(9):882–92. https://doi.org/10.1016/j.jcmg.2014.04.016.

[22] Montón Quesada I, Ogier AC, Ishida M, Takafuji M, Ito H, Sakuma H, Romanin L, Roy CW, Prša M, Richiardi J, Yerly J, Stuber M, Van Heeswijk RB. Self-gated free-running 5D whole-heart MRI using blind source separation for automated cardiac motion extraction. Magn Reson Med 2025;93(3):961–74. https://doi.org/10.1002/mrm.30322.

[23] Feng L, Coppo S, Piccini D, Yerly J, Lim RP, Masci PG, Stuber M, Sodickson DK, Otazo R. 5D whole-heart sparse MRI. Magn Reson Med 2018;79(2):826–38. https://doi.org/10.1002/mrm.26745.

[24] Schulz-Menger J, Bluemke DA, Bremerich J, Flamm SD, Fogel MA, Friedrich MG, Kim RJ, Von Knobelsdorff-Brenkenhoff F, Kramer CM, Pennell DJ, Plein S, Nagel E. Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update. J Cardiovasc Magn Reson 2020;22(1):19. https://doi.org/10.1186/s12968-020-00610-6.

[25] Prakken NH, Velthuis BK, Vonken EJJ, Mali WP, Cramer MJJ. Cardiac MRI: Standardized Right and Left Ventricular Quantification by Briefly Coaching Inexperienced Personnel. Open Magn Reson J 2008;1(1):104–11. https://doi.org/10.2174/1874769800801010104.

[26] McCarthy P. FSLeyes. Published online May 29, 2025. doi:10.5281/ZENODO.1470761.

[27] Ogier AC, Heskamp L, Michel CP, Fouré A, Bellemare M, Le Troter A, Heerschap A, Bendahan D. A novel segmentation framework dedicated to the follow-up of fat infiltration in individual muscles of patients with neuromuscular disorders. Magn Reson Med 2020;83(5):1825–36. https://doi.org/10.1002/mrm.28030.

[28] Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, Francis JM, Khanji MY, Lukaschuk E, Lee AM, Carapella V, Kim YJ, Leeson P, Piechnik SK, Neubauer S. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. J Cardiovasc Magn Reson 2016;19(1):18. https://doi.org/10.1186/s12968-017-0327-9.

[29] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11. https://doi.org/10.1038/s41592-020-01008-z.

[30] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, Jaeger PF. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv Prepr Post Online 2024. https://doi.org/10.48550/ARXIV.2404.09556.

[31] Ma J, Chen J, Ng M, Huang R, Li Y, Li C, Yang X, Martel AL. Loss odyssey in medical image segmentation. Med Image Anal 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035.
[32] Tustison NJ, Avants BB, Lin Z, Feng X, Cullen N, Mata JF, Flors L, Gee JC, Altes TA, Mugler, Iii JP, Qing K. Convolutional Neural Networks with Template-Based Data Augmentation for Functional Lung Image Quantification. Acad Radio 2019;26(3):412–23. https://doi.org/10.1016/j.acra.2018.08.003.
[33] Ogier AC, Hostin MA, Bellemare ME, Bendahan D. Overview of MR Image Segmentation Strategies in Neuromuscular Disorders. Front Neurol 2021;12:625308. https://doi.org/10.3389/fneur.2021.625308.
[34] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
[35] Bustamante M, Viola F, Engvall J, Carlhäll C, Ebbers T. Automatic Time-Resolved Cardiovascular Segmentation of 4D Flow MRI Using Deep Learning. J Magn Reson Imaging 2023;57(1). https://doi.org/10.1002/jmri.28241.
[36] Zange L, Muehlberg F, Blaszczyk E, Schwenke S, Traber J, Funk S, Schulz-Menger J. Quantification in cardiovascular magnetic resonance: agreement of software from three different vendors on assessment of left ventricular function, 2D flow and parametric mapping. J Cardiovasc Magn Reson 2019;21(1):12. https://doi.org/10.1186/s12968-019-0522-y.
[37] Ishida M, Yerly J, Ito H, Takafuji M, Nakamori S, Takase S, Ichiba Y, Komori Y, Dohi K, Piccini D, Bastiaansen JAM, Stuber M, Sakuma H. Optimal Protocol for Contrast-enhanced Free-running 5D Whole-heart Coronary MR Angiography at 3T. Magn Reson Med Sci MRMS J Jpn Soc Magn Reson Med 2024;23(2):225–37. https://doi.org/10.2463/mrms.tn.2022-0086.
[38] Roy CW, Di Sopra L, Whitehead KK, Piccini D, Yerly M, Heerfordt J, Ghosh RM, Fogel MA, Stuber M. Free-running cardiac and respiratory motion-resolved 5D whole-heart coronary cardiovascular magnetic resonance angiography in pediatric cardiac patients using ferumoxytol. J Cardiovasc Magn Reson 2022;24(1):39. https://doi.org/10.1186/s12968-022-00871-3.
[39] Sieber X, Romanin L, Bastiaansen JAM, Roy CW, Yerly J, Wenz D, Richiardi J, Stuber M, Van Heeswijk RB. A flexible framework for the design and optimization of water-excitation RF pulses using B-spline interpolation. Magn Reson Med 2025;93(5):1896–910. https://doi.org/10.1002/mrm.30390.
[40] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. *arXiv*. Preprint posted online. Image Is Worth 16×16 Words Transform Image Recognit Scale 2020. https://doi.org/10.48550/ARXIV.2010.11929.
[41] Zheng R, Wang Q, Lv S, Li C, Wang C, Chen W, Wang H. Automatic Liver Tumor Segmentation on Dynamic Contrast Enhanced MRI Using 4D Information: Deep Learning Model Based on 3D Convolution and Convolutional LSTM. IEEE Trans Med Imaging 2022;41(10):2965–76. https://doi.org/10.1109/TMI.2022.3175461.