DISSERTATION

A Multilevel Reader Comparison Software Tool for Semi-Automated Quality Control in Cardiovascular Magnetic Resonance Imaging: Lazy Luna

Eine mehrstufige Auswerter-Vergleichssoftware für halbautomatisierte Qualitätskontrolle in der kardiovaskulären Magnetresonanztomographie: Lazy Luna

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Thomas Hadler

Erstbetreuung: Prof. Dr. med. Jeanette Schulz-Menger

Datum der Promotion: 29.11.2024

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AHA	American Heart Association
AI	Artificial Intelligence
CMR	Cardiovascular Magnetic Resonance Imaging
CNN	Convolutional Neural Networks
CP	Clinical Parameter
CT	Computed Tomography
Dice	Dice Similarity Coefficient
DICOM	Digital Imaging and Communications in Medicine
EDV	End-Diastolic Volume
EF	Ejection Fraction
ESV	End-Systolic Volume
GUI	Graphical User Interface
HD	Hausdorff Distance
SAX	Short-Axis
LAX	Long-Axis
LGE	Late Gadolinium Enhancement
LL	Lazy Luna
LV	Left Ventricular
LVM	Left Myocardial Mass
ml diff	Millilitre Difference
PACS	Picture Archiving and Communications System
QA	Quality Assurance
QQ	Quantile-Quantile
RV	Right Ventricular
SV	Stroke Volume

Abstract

Cardiovascular magnetic resonance imaging (CMR) provides a non-invasive and detailed assessment of cardiac structures and tissues, offering valuable quantitative assessments for diverse cardiac conditions. Quantification requires precise annotations by readers, including reference points and delineations of heart chambers, myocardium and other structures. As clinical applications multiply and readers are supported by the clinical integration of artificial intelligences (AI), evaluating the reproducibility of readers and quantification methods becomes essential. The aim on this thesis is to design, implement and test an extendible software tool, Lazy Luna (LL), that is dedicated to semi-automated multilevel reader comparison to increase quality control in CMR. A multilevel comparison of readers offers annotation comparisons, quantitative clinical parameter (CP) comparisons, and causal explanations for CP differences with annotation differences.

First, a software prototype was designed and implemented for short-axis (SAX) cine imaging. Annotations were modelled as geometric objects, ensuring exact calculations of segmentation metric values and CPs. Difference tracing was implemented to allow for finding causal annotation explanations for CP differences between readers. A graphical user interface (GUI) was implemented to enable user-inspection of statistical reader differences and locating their origins in annotations. LL was tested by comparing two CMR experts to each other who annotated 13 SAX cine datasets. Second, interfaces were implemented for all software components (e.g. CPs, visualizations, tables) to facilitate the extendibility of LL to new imaging sequences. The extendibility was tested by applying LL to 13 parametric T1 mapping cases annotated by two expert readers and Late Gadolinium Enhancement (LGE) cases from the openly available Emidec dataset with reference contours and AI contours.

First, the software prototype of LL calculated precise segmentation metrics and CPs. The GUI allowed for tracing large CP differences to contouring differences. The SAX reader comparison revealed that differences in basal slices contour choices led to large volumetric differences. Second, the extendible back- and frontend of LL allowed for extensions to T1 mapping and LGE. For T1 mapping LL was used to trace CP differences to contouring variability. For LGE LL revealed an undertrained AI that had learned the myocardial contour, but poorly estimated scar tissue. The application cases showed that LL is useful open-source software for reader comparison on multiple imaging techniques.

The semi-automated multilevel reader comparison software, Lazy Luna, was successfully implemented for several CMR imaging techniques. LL calculates differences for large cohorts and offers qualitative explanations for biases between readers. LL provides insights into the challenges of CMR standardization and promising technologies for the future of quantitative CMR.

Zusammenfassung

Die kardiovaskuläre Magnetresonanztomografie (CMR) ermöglicht eine nicht-invasive Funktions- und Gewebeanalyse, die quantitative Werte zur kardialen Diagnostik beiträgt. CMR Quantifizierung erfordert Annotationen durch Auswerter, mit Referenzpunkten und Konturierungen von Herzkammern, Myokard und weiteren Strukturen. Zunehmende Anwendungen und die klinische Integration künstlicher Intelligenzen (AI) benötigen die Evaluation der Reproduzierbarkeit von Auswerter und Quantifizierungsmethoden. Ziel dieser Arbeit ist das Entwerfen, Implementieren und Testen eines erweiterbaren Software-Tools, Lazy Luna (LL), das dem halbautomatisierten mehrstufigen Auswertervergleich zur Qualitätskontrolle in der CMR gewidmet ist. Ein mehrstufiger Vergleich von Auswertern bietet Annotationsvergleiche, klinische Parameter (CP)-Vergleiche und Annotationsgründe für CP-Unterschiede.

Zuerst wurde ein Software-Prototyp für die Kurzachsen-cine (SAX)-Bildgebung entworfen und implementiert. Annotationen wurden geometrisch modelliert um exakte Berechnungen von Metriken und CPs zu gewährleisten. Die Verfolgung der Differenzen erlaubte es Annotationserklärungen für CP-Unterschiede zwischen Auswertern zu ermöglichen. Eine grafische Benutzeroberfläche (GUI) wurde implementiert, um die Nutzerinspektion von statistischen Auswerterunterschieden zu ermöglichen und ihre Annotationsursprünge zu ermitteln. LL wurde getestet, indem zwei Experten miteinander verglichen wurden, die 13 SAX-Cine-Datensätze annotierten. Zweitens wurden Schnittstellen für alle Softwarekomponenten (z.B. Visualisierungen, Tabellen) implementiert, um die Erweiterbarkeit von LL auf neue Bildsequenzen zu erleichtern. Die Erweiterungen wurden getestet, indem LL zum Vergleich von zwei Experten auf 13 parametrische T1 Mappingbildern, und von Referenz- und AI-Konturen auf Late Gadolinium Enhancement (LGE) Bildern angewendet wurde.

Zuerst berechnete der Software-Prototyp präzise Metriken und CPs. Die GUI ermöglichte das Verfolgen von CP-Unterschiede zu ursächlichen Konturierungsunterschieden. Der SAX- Auswertervergleich zeigte Konturunterschiede die in basalen Schichten zu großen Volumenunterschieden führten. Zweitens ermöglichte die erweiterbare Back- und Frontend von LL Erweiterungen auf T1-Mapping und LGE. Für T1-Mapping wurde LL verwendet um Konturierungsunterschiede für CP-Unterschiede zu finden. Für LGE offenbarte LL eine AI, welche die Myokardkontur gelernt hatte, aber Narbengewebe unterschätzte. Die Anwendungsfälle zeigten, dass LL eine nützliche Open-Source-Software für den Auswertervergleich mehrerer CMR Bildgebungstechniken ist.

Die halbautomatisierte mehrstufige Auswertervergleichssoftware, Lazy Luna, wurde erfolgreich für verschiedene CMR-Bildgebungstechniken implementiert. LL berechnet Auswerterunterschiede und bietet qualitative Erklärungen für dieselben. LL ermöglicht Einblicke in die Herausforderungen der CMR-Standardisierung und vielversprechende Technologien für die Zukunft der quantitativen CMR.

1 Introduction

1.1 Cardiovascular Magnetic Resonance Imaging

Cardiovascular magnetic resonance imaging (CMR) is a non-invasive medical imaging technique that has transformed the assessment, diagnosis, and treatment of various cardiovascular diseases [1–3]. CMR exploits the heart's magnetic properties to provide high-resolution images of the heart and blood vessels, allowing healthcare professionals and researchers to gain invaluable insights into the structure and function of the cardiovascular system [1,2,4]. CMR has become an essential tool in cardiology, offering superior diagnostic accuracy, a wealth of information for treatment planning, and is becoming ever more prevalent and accepted in guidelines and recommendations, such as in the European Society of Cardiology [5,6].

CMR offers several advantages over other imaging modalities, such as computed tomography (CT) and echocardiography. Unlike X-Ray or CT, CMR is void of ionizing radiation exposure, making it safer for recurring examinations, and extends the diagnostic toolkit beyond echocardiography [4]. In addition to the capability of CMR to assess cardiac structures from multiple angles, it also assesses surrounding tissues, the pericardium, great vessels, and possible masses in the chest [3]. Further, CMR provides cardiac function information such as ventricular outputs, ejection fractions, and blood flow as well as myocardial tissue composition with parametric mapping, late gadolinium enhancement and perfusion imaging. This allows CMR to diagnose complex cardiovascular conditions such as congenital heart diseases, and cardiomyopathies, as it enables precise characterization of tissue properties and blood flow dynamics [5].

1.1.1 Imaging Techniques

Cine imaging provides high space-and-time resolved cardiac images for volume and function assessment of the ventricles and atria as well as revealing wall-motion behaviour [3]. Short-axis (SAX) cine imaging covers the left and right ventricles (LV, RV) with a stack of slices perpendicular to the septal wall. Long-axis (LAX) cine imaging positions slices to show all four heart chambers. Cine imaging is crucial for diagnosing heart failure, cardiomyopathies and valvular diseases. Parametric mapping techniques allow for tissue characterization by estimating underlying bio-physical properties as numerical values, which are assigned to voxels [7]. The parametric approaches allow for quantitative myocardial tissue differentiation; T1 mapping allows for differentiating healthy myocardium from diffuse fibrosis and detecting infiltrative diseases (e.g. amyloidosis or fat accumulation), T2 mapping reflects the myocardial water content, providing oedema and inflammation imaging [8–11]. Late gadolinium enhancement (LGE) employs a contrast agent, which accumulates in scarred tissue and allows for focal scar and fibrosis detection as well as quantification [12,13]. Phase-contrast CMR encodes the speed and direction of blood flow patterns, which allows for the assessment of hemodynamics [14]. CMR can combine these imaging techniques within a single examination, making it a versatile, non-invasive tool for characterizing and diagnosing a wide range of cardiovascular conditions [5,6].

1.1.2 The Imaging Chain

The CMR imaging chain starts with patient preparation (e.g. obtaining the medical history, assessing illnesses and medical implants), followed by their positioning on the examination table where radiofrequency coils are placed over their chest [1,2]. Before patient-specific images are acquired localizer scans are performed to determine imaging planes and regions of interest [3]. By selecting different pulse sequences (“imaging techniques”) raw data are acquired. To minimize motion artefacts, pulse sequences are synchronized to the cardiac cycle with cardiac gating, which ensures the data are acquired at specific phases of the heartbeat. In addition to this, breath-hold commands may minimize respiratory motion artefacts. The acquired raw data is then reconstructed into images; for example, during cine imaging a sequence of images is captured throughout the cardiac cycle. Following the reconstruction, post-processing steps may remove motion artefacts and reduce noise [15]. These final CMR images are visually analysed and annotated by a CMR expert, who judges whether the image quality is acceptable for interpretation, and then pinpoints and delineates relevant cardiac structures visible within the images [3,16]. Based on the images and their annotations clinical parameters pertaining to cardiac function, myocardial tissue characteristics and quantifications, and abnormalities are calculated and remarked upon. The findings are collected in a report, including measurements, clinical remarks and diagnostic recommendations, which are integrated into the patient’s overall clinical evaluation.

1.2 Confounders in the Imaging Chain

Although it is generally agreed that CMR imaging techniques offer a high accuracy and reproducibility, they are affected by a multitude of confounders in the CMR imaging chain [17]. Variability-inducing confounders can originate in all steps of the imaging chain, from different scanner sites, image acquisition and reconstruction methods, image annotation and clinical parameter assessment, to patient diagnosis and prognosis. Different groups use different vendors, scanners, sequences, and technicians, who may position coils and perform the image planning and positioning during the scan. Depending on the overall system, different imaging techniques may be employed to acquire the same image types, for example parametric T1 mapping can be performed with several different pulse–sequences [10]. Patients may have atypical cardiac morphologies, making default imaging planes problematic, or have difficulties breathing, which can produce motion artefacts and noise during image acquisition or perturb the image reconstruction and post-processing of images. Furthermore, cardiac gating is known for its vulnerability [18], may impact image quality and lead to unreliable quantification. The image analysis and annotation may suffer from reader variability, which can impact annotation reproducibility and affect clinical parameters and diagnostic decisions. Reader variability may further affect the reports of CMR experts, when standardised reporting fails to mitigate reader differences.

In order to decrease the effects of confounders, standardisation initiatives are continuously investigated. Such initiatives include multi-vendor, multi-site [19] and travelling volunteer studies [20], which are used to assess the comparability of parameters obtained from different sites, while comparability methods, like the Z-score [21] intend to correct differences between sites or scanners. Imaging protocols are published in order to standardise and streamline patient-based CMR to decrease variability caused by imaging procedure [22]. Innovations in image reconstruction, such as parallel MRI reconstruction and compressed sensing have greatly impacted scanning times and thus the set of clinically applicable sequences [15,23,24]. However, reconstruction algorithms can be prone to producing artefacts, and difficult to reproduce due to unavailable source-code. Open-source frameworks, such as Gadgetron intend to ameliorate this by providing comprehensive, and standardised approaches to image reconstruction [25,26]. Visual image quality guidelines and image/signal-to-noise-ratio classification algorithms aim to provide reproducible image quality assessments for artefact detection and data exclusion [22,27]. Quantitative imaging biomarkers provided by radiomics, machine learning, and convolutional neural networks (CNN) offer information about physiological structures in the image [28–30]. However, for radiomics algorithms source-code is not always available, and for CNNs the variety of training datasets and procedures can lead to difficult-to-understand blackboxes. To address these drawbacks the “Image Biomarker Standardisation Initiative” offers reproducible radiomics features [31]. In order to further streamline the post-processing of CMR images post-processing guidelines are employed [3]. For artificial intelligences (AI), public segmentation competitions provide benchmarking, reproducible results and open-source code [32–34]. Although significant efforts have been made to reduce the effects of confounders on the CMR imaging chain, many sources of variability remain insufficiently mitigated.

1.3 Reader Differences for Segmentations and Clinical Parameters

Even though guidelines and standardising consensus statements are well established in the CMR community, reproducibility in intra- and inter observer studies showed that significant annotation and clinical parameter differences remain [20,35]. For SAX cine, highly variable annotation decisions for basal slices cause large volumetric differences, while apical slices remain similarly irreproducible with smaller volumetric impacts [36,37]. This may be caused by different annotation decisions, fat in the images, or by partial volume effects, in which single image voxels contain a mixture of different tissues or cardiac structures due to their large slice-thicknesses (7–10mm) compared to their pixel spacings. And although clinical parameter differences are close to zero for interobserver assessments, the variance can be high even for expert readers. In 2015, Suinesiaputra et al. demonstrated the necessity of further standardisation by comparing seven experts to each other, who each contoured 15 cases according to SCMR guidelines [38]. Large differences in segmentation decisions for LV basal and apical slices and clinical parameters were revealed between sites. For parametric mapping, apical slice value assessments are less reproducible than those for basal or midventricular slices [39]. And a recent study on the reproducibility of T1 and T2 parametric mapping parameters across sites and readers, implied that mapping parameter differences between readers also explain some of the variability between sites [40]. Since inter- and intraobsever analyses isolate reader-specific contributions to segmentation and parameter differences, and differences still remain in intraobserver analyses, some relevant segmentation differences must be due to the difficulty of image interpretation.

Training CMR newcomers allows for mitigating segmentation differences and increasing the reproducibility of clinical parameters in-site [41], while simultaneously spreading the post-processing guidelines, which offers additional standardisation across sites and countries. Typically, trainees have a medical background with vast knowledge in cardiology and experience in medical imaging. For post-processing of CMR images, educating trainees builds on curriculums and face-to-face teaching with immediate feedback. Training has been shown to increase the reproducibility of LV volume assessments [42]. Following courses on the basics of image interpretation, supervised practice in clinical routine leads to increased familiarity and responsibility of newcomers with CMR. However, as CMR availability grows around the globe, the clarification and imparting of post-processing techniques and guidelines grows more difficult and expensive in turn [43,44].

AI tools provide fully or semi-automated segmentation algorithms for cardiac structures, calculating clinical parameters with errors in the order of interobserver variability [32,33,45]. AI tools address several challenges of manual segmentation, which can be time-consuming and prone to interobserver variability. They are employed in research to analyse large datasets of CMR images or to speed up manual segmentation tasks in clinical routine, where CMR experts check AI outputs. In addition to clinical parameter evaluations, segmentation CNNs are trained and evaluated with segmentation metrics, such as the Dice similarity coefficient (Dice) [46] and the Hausdorff distance (HD) [47]. However, while CNNs achieve high segmentation metric values similar to experts, they continue to produce human-atypical segmentation failures, such as fragmented segmentations that violate cardiac geometry constraints (e.g. myocardial contours that run through the bloodpool) and cause distrust in AIs [48–50]. CNNs promise a higher efficiency, and scale indefinitely, providing new hope for widespread standardisation that could be shared across sites, this promise of standardisation is doubtful due to the ever-expanding number of available AIs. Most post-processing vendors offer supporting AI segmentations for several CMR imaging techniques, and segmentation competitions have produced many more. The comparability of these algorithms is unclear and the varying results, so far attained in diverse competitions, imply that the human reader variability may instead be reproduced with AIs.

To summarise the nature of reader differences: segmentation standardisation and segmentation improvement can be opposing goals. Consensus and guidelines make the point that standardising procedures ought to lead to more reproducible image segmentations and clinical parameters [3,9]. At the same time consensuses and guidelines tend to acknowledge that significant variability remains due to lacking agreement and genuine difficulties, unknown confounders, and also acknowledging that certain inconsistencies may be justified due to the evolving methodologies, such as the clinical integration of AIs or the increased differentiation of cardiac structures [35,51].

1.4 The Need for Quality Control

Quality assurance has been pursued along the CMR imaging chain in order to minimise the effects of confounders. As illustrated above, to standardise reader segmentation of images, guidelines and training programs were introduced. Research tends to focus on reader comparisons to judge the reproducibility of clinical parameter assessments of a multitude of imaging techniques within and across sites. The methodological focus of clinicians and AI developers on reader comparison should be integrated into an overall comparison method, called a multilevel reader comparison. On the segmentation level, clinicians tend to focus on qualitative differences, including segmentation difference visualisations, descriptions of cardiac structures and how guideline-consistent segmentations behave in relation to them. AI developers tend to focus on quantitative segmentation metrics to assess segmentation proficiency. Multilevel reader comparison should assess and explain biases between two readers as statistical properties of the assessed clinical parameters. The assessed clinical parameter differences should in turn be explained by segmentation differences – described quantitatively and qualitatively. In practice, multilevel reader comparisons would be tedious tasks, full of laborious and error-prone subtasks, including clinical parameter calculations, statistical calculations, assessments of statistical clinical parameter acceptability, and visualisations for a better understanding of reader similarities and deviations, while allowing to trace reader differences to individual clinical parameter differences back to their causal segmentation difference origins. A semi-automated quality control tool that performs a multilevel reader comparison could be readily applied to assess the acceptability of clinical parameter assessments in new environments. Currently, the CMR landscape of software tools is missing a reader comparison tool to automate and streamline such a multilevel reader comparison.

1.5 Aims

The overall aim of this thesis is to design and implement an extendible semi-automated multilevel reader comparison tool for quality assurance tasks in cardiovascular magnetic resonance imaging. The thesis focuses on two publications, which pertain to comparison software, named Lazy Luna (LL). The first publication aims to develop a software prototype (“Software Prototype Development”), which illustrates the general functionality of such a tool on short-axis cine images [52]. The second publication aims to formally design LL as generalizable and extendible software (“Software Architecture Design”), which ensures the software can be extended to an ever-changing environment of imaging techniques, clinical parameters and confounders in CMR [53].

2 Methods

2.1 Outline of Software for Multilevel Reader Comparison

The software Lazy Luna was designed to offer a multilevel reader comparison that covers segmentations, clinical parameters and reader statistics (Figure 1). CMR images, segmentations and clinical parameters were modeled to allow for accurate calculations of segmentation metrics, clinical parameters and reader statistics. These analysis levels were connected to allow for tracing differences from the segmentation level over the parameter level to the statistical reader level (e.g. biases). An interactive GUI was designed to make the software accessible.

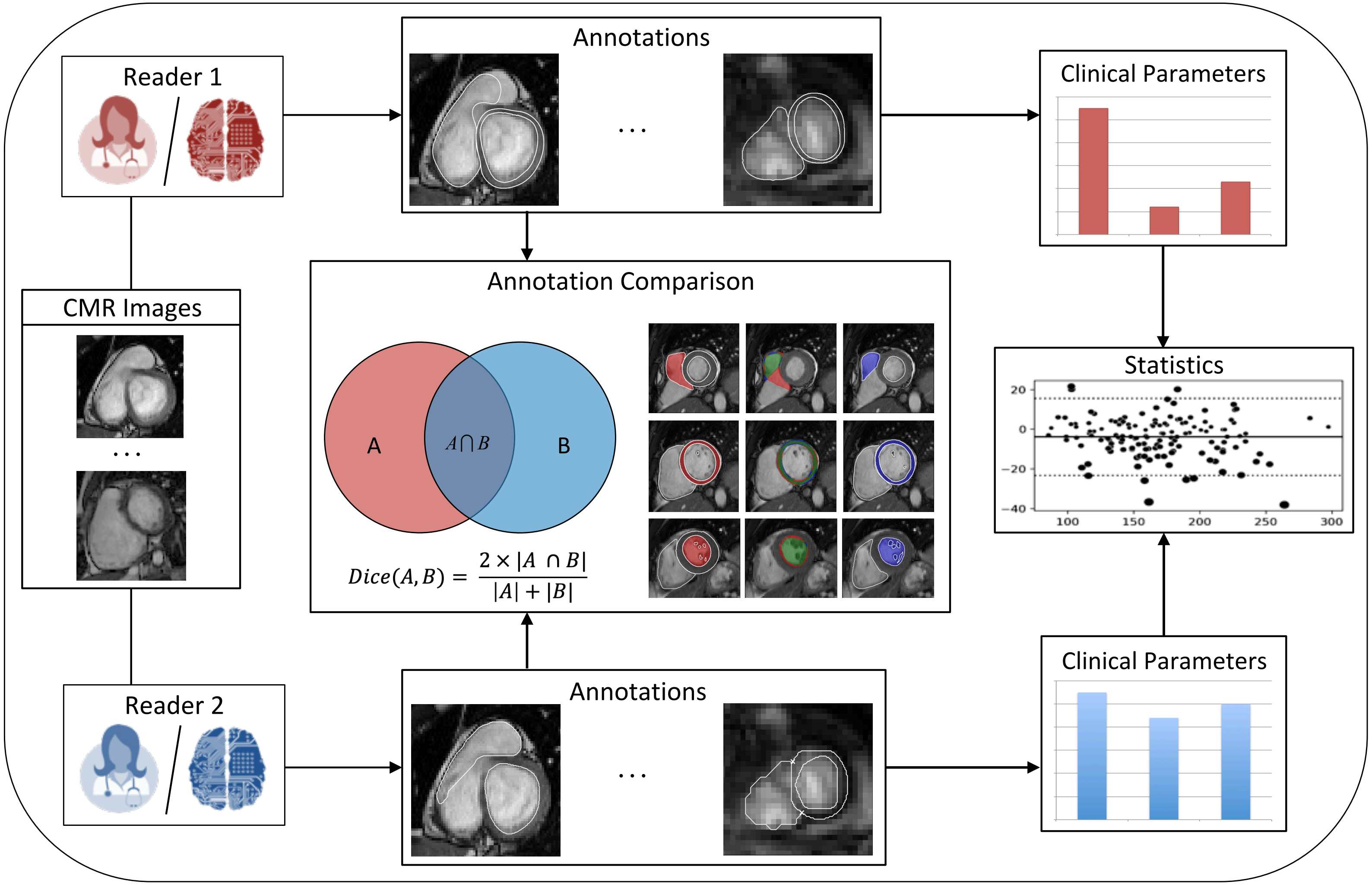


Figure 1: Multilevel Reader Comparison

Attribution: adapted from "Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging" by Hadler et al. 2022, https://www.nature.com/articles/s41598-022-10464-w, Licensed under a Creative Commons Attribution 4.0 License.

Caption: Cases can be compared when two readers annotated the same CMR images (first reader above in red, second below in blue). Clinical results are calculated from images and annotations. Image segmentations can be compared with segmentation metrics such as the Dice similarity coefficient (Dice) or the Hausdorff distance. Segmentation differences can be visualized as overlapping or disjoint by color-coding them (central subfigure). Reader biases can be visualized as statistical plots, such as Bland-Altman plots (right) of reader differences of cardiac volume assessments.

2.2 Software Prototype Development

This subchapter describes several focuses during LL's development. This included the data interface, how quantification accuracy was guaranteed, how difference tracing from annotations to clinical parameter impacts was implemented, and how the usability of the software was tested.

2.2.1 Data Interface

Images

The DICOM [54,55] format (Digital Imaging and Communications in Medicine) is the international standard for the medical imaging communication. Keeping data in this format avoids conversion loss for images and adjacent information, which is relevant for sorting images and annotations, visualisations and parameter calculations, including: a unique identifier, acquisition time stamp, image position in 3D space, pixel spacing, slice thickness, and pixel value conversion parameters.

Annotations and Geometrical Operations

Annotations consist of geometric objects (e.g. points, lines, polygons and multi-polygons), which offer accurate representations of sub-pixel annotations of cardiac structures (e.g. insertion points, ventricular cavities, papillary muscles) by expert readers in post-processing software, as well as precise delineations of pixel masks (typical AI outputs). Geometric annotations allow for precise geometrical operations, including intersections, unions, distance and area calculations.

2.2.2 Quantification Accuracy

Segmentation Metrics

The geometrical operations for annotations allow for calculating segmentation metrics, such as Dice and HD. In addition to this, DICOM attributes such as the pixel width, height and slice thickness allow for calculating the volume differences in millilitres (ml diff). Dice measures the overlap of two polygons (areas A and B); HD measures the maximum among the minimal distances between the outlines of two polygons (contours of A, B: cA, cB). The ml diff provides the volumetric impact of a segmentation difference.

$$\mathrm{Dice}(A,B)= \frac{2 \times |A \cap B|}{|A|+|B|}$$

$$\mathrm{HD}(A,B) = \max\left\{ \max_{a\in cA}\left(\min_{b\in cB} d(a,b)\right),\ \max_{b\in cB}\left(\min_{a\in cA} d(a,b)\right) \right\}$$

$$\text{ml diff}(A,B)= (|A|-|B|) \times \text{area per pixel} \times \text{slice thickness}$$

Clinical Parameters

Short-axis clinical parameters include ventricular volumes, such as end-systolic and end-diastolic volumes (ESV, EDV) for the left and right ventricle (LV, RV) and the LV myocardial mass (LVM). Cardiac function parameters include the LV and RV stroke volume (SV) and ejection fraction (EF). All parameters are based on volume calculations, which are calculated on the basis of annotations, pixel height and width, and slice thickness (Figure 1). For each parameter a class was implemented to simplify calculation and comparison.

Cases

Cases were implemented as container classes (i.e. classes that contain and organize several interacting objects) in order to facilitate the connection between images, annotations and clinical parameters as well as connect the multiple levels of analysis to each other. A case contains references to all images from one CMR-scan, image annotations from one reader, and clinical parameters calculated from the images and annotations.

2.2.3 Difference Tracing

Difference tracing required visualisations that compare annotations on the image level and readers statistically so that differences between readers could be understood qualitatively and quantitatively.

Annotation Comparison Visualisation

A visualisation of annotation differences was implemented by calculating intersecting and disjoint surfaces to plot them in separate colours (Figure 1). This shows regions that were exclusively segmented by the first or the second reader, and agreement between both.

Statistical Reader Comparison Visualisation

When several cases were annotated by two readers they can be compared to each other (Figure 1). Quantile-quantile plots, Paired Boxplots and Bland Altman plots are implemented as visualisations of statistical clinical parameter reader differences. These plots include statistical information: quantile-quantile plots are used to visually estimate whether a Gaussian distribution may be assumed, paired boxplots show a quantile boxplot for both readers, with all cases' values scattered as points on top of the boxplots. As each case has a specific clinical parameter value, as assessed by both readers, the points can be connected by a line when they represent the same case. Bland Altman plots represent the clinical parameter comparisons as the average on the x-axis and the difference between both assessments of the y-axis.

Implementation of Difference Tracing

Difference tracing was implemented by connecting point elements of the statistical visualisations to the cases. Since the statistical visualisations "know" which cases are represented by the individual points, they can communicate which case was selected to the program. By clicking on a point in a statistical visualisation, another tab is opened that focuses on the selected case as annotated by both readers. This tab focuses on annotation comparison for the case's images, with metric value calculations and a visualisation of the annotations for both readers.

2.2.4 Usability

Software Package Dependencies

LL was written in Python 3.8, and requires several software packages to run:

- Pydicom [56] 2.2.0
- Shapely [57] 2.0.0
- Pandas [58] 1.2.4
- Matplotlib [59] 3.6.2
- Seaborn [60] 0.11.1
- PyQt5 [61] 5.15.7

User interface

A graphical user interface (GUI) was written in PyQt5. The user can select cases from two readers and compare them to each other statistically in a tab. The interactive matplotlib figures allow for opening another tab in the GUI, in which individual annotations are compared. Tabular information can be exported for the user.

Software Prototype Testing

The prototype was verified by performing a multilevel interobserver analysis on short-axis images stacks of 13 scans from a 1.5T Avanto fit (Siemens Healthineers, Erlangen, Germany). These images were annotated by two readers in cvi42 (Version 5.12.1, Circle Cardiovascular Imaging, Calgary, Canada) to produce segmentations on LV, RV endocardia, as well as LV papillary muscle and myocardium. The two readers were compared with LL.

LL was used to explain reader volume differences with annotation differences, and exported tabular data. First, LL was used to show the impact of segmentation differences on clinical parameters by isolating segmentation metric values for contour comparisons that affected specific parameter differences and presenting the results in a table. Second, LL was used to analyse where segmentation errors originated in cardiac geometry for each contour type and presented this tabular information.

2.3 Software Architecture Design

Building on the software prototype, LL was formally designed as generic software that is extendible to the evolving requirements of CMR, such as new imaging techniques and clinical parameters. This necessitated a documentation of formal software requirements.

2.3.1 Requirements

Accessibility and product independence

LL should be independent of vendors or reader output (polygons for human readers, image masks for CNNs) and should be available as open-source software building on open-source components.

Extendibility

The following is adapted from Hadler et al. 2022 [53]: “LL requires an understandable backend so that developers can extend the software to new sequences. LL’s core components should” be implemented as “classes, to allow for a generic, extendable backend.

Usability and Target Groups

LL should be usable by” AI “developers and medical experts. A graphical user interface (GUI) should be provided in order to allow for an automatic reader comparison” for multiple imaging techniques with “expressive visualizations and statistical analyses, independent of programming familiarity.”

2.3.2 Accessibility and Product Independence

Generic Interface for Annotations

A custom LL annotation format stores geometric representations of annotations as dictionaries with key-value pairs, connecting annotation types (e.g. contour or point name) to their annotations (i.e. Shapely geometry and auxiliary information). The dictionaries were stored in pickle format with the associated DICOM’s unique identifier as its basename, the SOPInstanceUID.

Open Source

LL should be made available as open-source software on Github.

Product Independence and Code Maintenance

LL should be designed with open-source packages for its functionalities. LL requires a maintenance plan since the packages evolve, including code and API, which may perturb LL’s usability. This dissertation’s author maintains LL’s open-source code.

2.3.3 Usability and Target Groups

Agile Development Procedure

For the prototype, clinicians contributed information on relevant clinical parameters and preferred visualisations of annotation differences. LL required a development procedure that integrated clinician feedback in iterative feedback loops for incremental improvement (Figure 3).

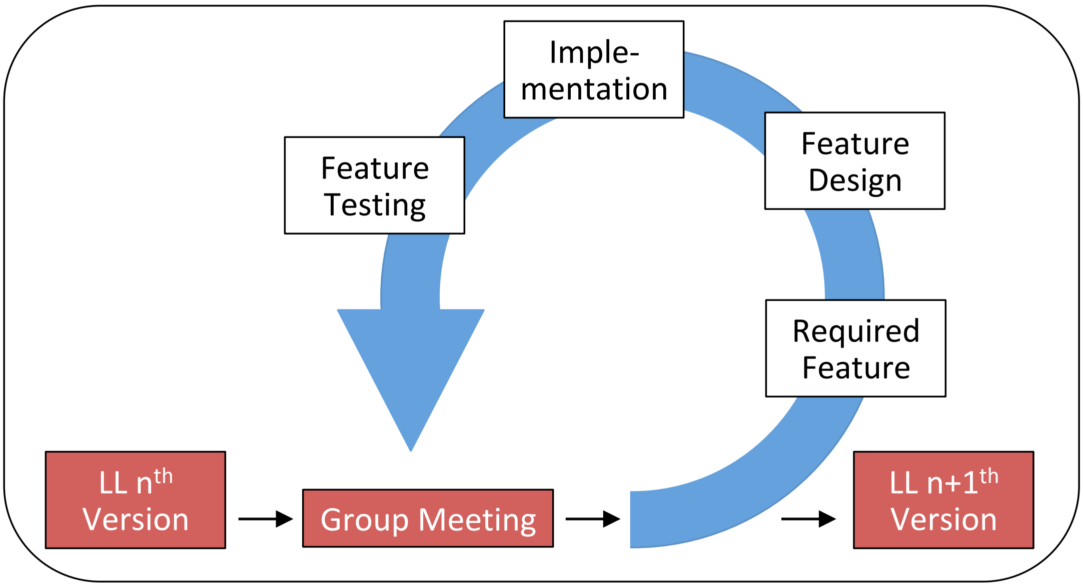


Figure 2: Agile Development Procedure of Lazy Luna
Attribution: adapted from "[Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging]" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.
Caption: Starting from Lazy Luna's current state (n), a development cycle consisted of user meetings who described desired functionalities, followed by abstractions thereof to implementable features, their implementation and testing These were then discarded or integrated, depending on clinician/AI developer opinion during the next group meeting. Following such a development cycle, Lazy Luna version n+1 was installed.
Legend: LL: Lazy Luna

Graphical User Interface

The GUI described in "Software Prototype Development" (2.2) was redesigned to build on the formalized classes and offer the selection of different views for each available imaging technique.

2.3.4 Extendibility

Class Diagram

In order to make LL understandable and easily extendible it follows an object-oriented programming paradigm. LL's class structure is presented in Figure 3.

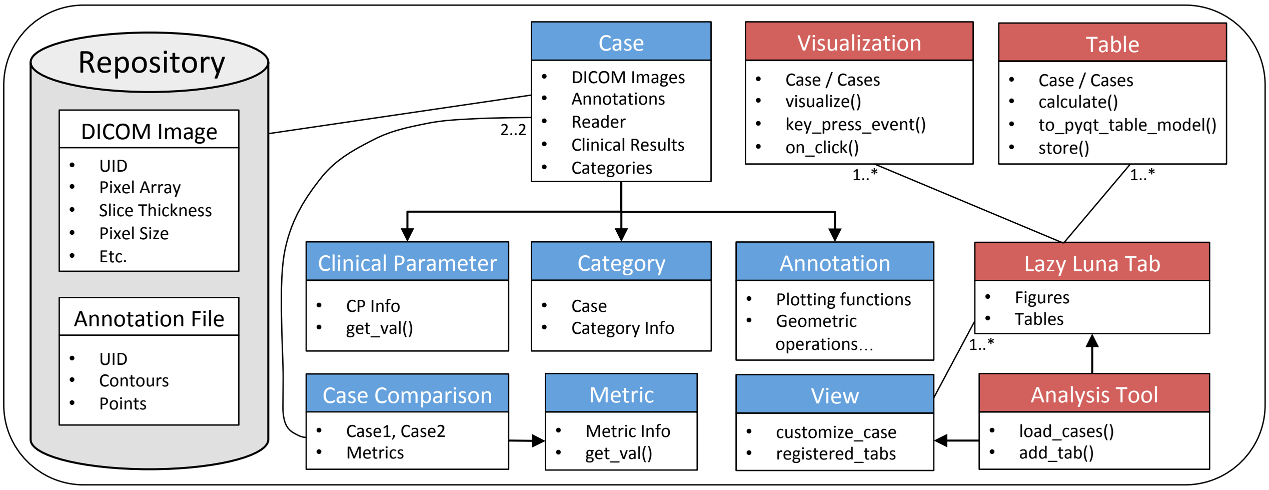


Figure 3: Lazy Luna Class Diagram
Attribution: adapted from "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.
Caption: The data repository (grey) contains DICOM images and annotation files. The software backend (blue) shows backend classes and relationships. Cases contain DICOM image and annotation references, a reader, clinical results and categories. (Classes in capital letters) Clinical Results calculate clinical parameters for the case. Categories can be attached to Cases for simplified access to DICOMS and annotations. Metrics compare two annotations to each other quantitatively. View classes organize the other classes to address use-case needs. The frontend (red) consists of Visualizations and Tables as GUI elements as well as Tabs (PyQt5 Widgets) and the Analysis Tool, the main application interface.
Legend: DICOM: digital imaging and communications in medicine, GUI: graphical user interface

Extending Classes to New Imaging Techniques

LL should be extendible to other imaging techniques by extending the backend classes in Figure 2. Implementing new Clinical Parameter classes or Metric classes may require code extensions of the Annotation or Category classes. LL's extendible classes, their easy access and modification functions (getter and setter functions) as well as extension options are presented in Table 1. Annotation is a utility class that handles pickled Python dictionaries containing Shapely geometries. New geometrical operations or visualisations can be implemented as Annotation class functions. Category is a class that sorts images and annotations. Adjusting LL to other imaging techniques may require the implementation of additional clinical parameter classes. Metric classes can be added to quantify annotation similarity. View classes organise cases to address user needs for an imaging modality. When new tabs are designed for an imaging modality, they are registered in a View class.

Table 1: **Class Descriptions and Extendibility**

Attribution: adapted from "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.

Class	Use Description	Functions	Extension Description
Annotation	Interface class to geometries: - Getter functions: provide access to geometry objects - Visualizations: of geometries atop matplotlib axes - Helper functions: offer complex geometrical calculations	Getter functions: - get_contour(cname) - get_point(pname) Visualization functions: - plot_contours(axis, cname, color) - plot_points(axis, pname, color) - plot_face(axis, cname, color) - plot_cont_comparison(axis, other_anno, cname, colors) Helper functions: - get_contour_as_mask(cname)	How to: The class is extended by adding new functions Exemplary helper functions: - Point distances - Angle calculations - Bounding box determination
Category	Sorts images and annotations: - Sorting functions: sort images and annotations spatially and temporally according to DICOM attributes - Getter functions: provide access to DICOMs, images and annotations - Helper functions: offer calculations that require sorted DICOMs and annotations	Sorting function: - get_sop2depthand-time(sop_uid2filepath) Getter functions: - get_dcm(slice, phase) - get_img(slice, phase) - get_anno(slice, phase) Helper functions: - get_volume(cont_name, phase)	How to: The class is extended by adding new functions Exemplary helper functions: - Cardiac geometry descriptions (such as basal, midventricular, apical slices) - Determining phases in cardiac cycle (like end-systolic phase)
Clinical Result	A Clinical Result calculates a clinical parameters for a case	Setter functions: - init(case): sets case, clinical parameter name, measurement unit Getter functions:	How to: Lazy Luna is extended by clinical results for new imaging techniques by writing new classes

		- get_val(as_string=False) - get_val_diff(other_clinical_result, as_string=False)	Exemplary extension: - Clinical result for calculation of average myocardial voxel intensity
Metric	A Metric quantifies the difference between two annotation geometries with the support of corresponding DICOMs Metric values: calculation of metric values	Setter functions: - init(): sets metric name, measurement unit Getter functions: - get_val(geo1, geo2, dcm=None, as_string=False)	How to: Lazy Luna is extended by metrics for new imaging techniques by writing new classes Exemplary extension: - Difference in number of pixels within contour type for two readers
View	A View structures cases by appending relevant categories, clinical results and tabs	Setter functions: - init(): sets the view name, tabs for inidividual case_comparisons and lists of case comparisons Adjust case: - initialize_case(case): calculates information requiring only one calculation - customize_case(case): connects the view's categories and clinical results Store information function: - store(case_comparisons)	How to: Extending Lazy Luna to a new imaging modality requires the implementation of a custom View class Exemplary extension: - A View for focal scar imaging

Plug-in Scheme

LL requires a generic plug-in scheme to allow for simple visualisation and table element integration into the GUI. Figures inherit their functionality from the matplotlib.Figure class and can be integrated into PyQt5.QWidget [62] objects with the FigureCanvas [63] interface class offered by matplotlib. Tables function as an interface class allowing for seamless transformation of Pandas DataFrame [64] objects to the QtGui.QStandardItemModel for table presentation.

2.3.5 Extension to T1 Parametric Mapping & Late Gadolinium Enhancement

In order to verify the extendibility of Lazy Luna, the software was extended to two imaging techniques. First, LL was extended to parametric T1 mapping and tested on a dataset of 13 parametric T1 mapping cases, which were annotated by two clinicians with cvi42 (Version 5.12.1, Circle Cardiovascular Imaging, Calgary, Canada). Second, LL was extended to accommodate LGE imaging. To this end LL is tested on the openly available Emidec dataset [34] by comparing the publicly available segmentation masks provided as reference masks to an AI's segmentation masks.

3 Results

3.1 Software Prototype Development

3.1.1 Data Interface

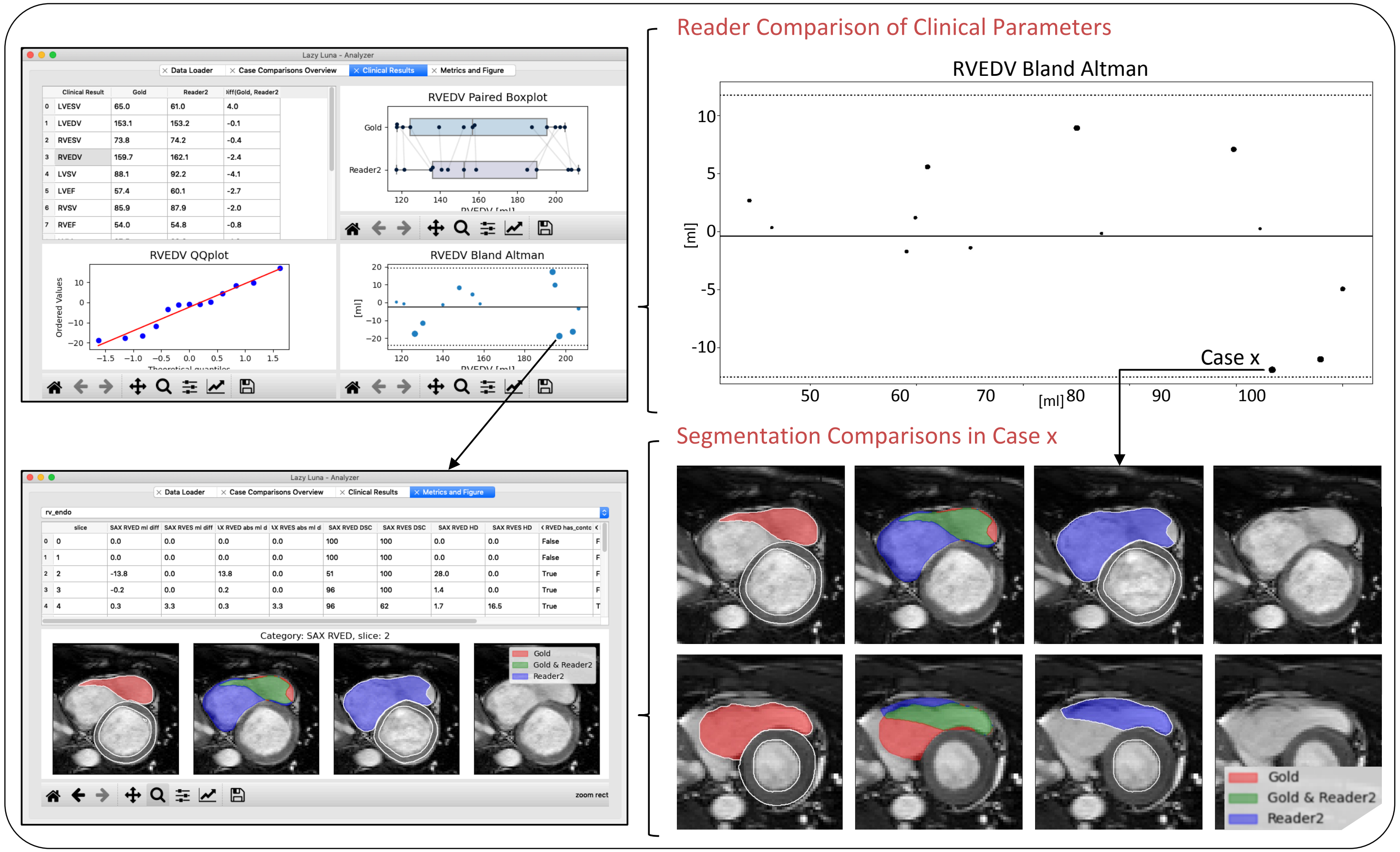
Images were stored in DICOM format, which guaranteed complete image information. Annotations were stored as geometrical representations, which ensured precise segmentation delineations regardless of reader type (human or AI).

3.1.2 Quantification Accuracy

Segmentation metric and clinical parameter calculations were based on exact geometrical representations, which guaranteed sub-pixel accuracy, and DICOM tags, which represent the image information precisely as output by the scanner.

3.1.3 Difference Tracing

Offered within the GUI, difference tracing was implemented to track clinical parameter biases to annotation differences. This is demonstrated in Figure 4, where an LL user searched for reasons for larger reader deviations of the RV EDV, visible within a Bland-Altman plot. This qualitative investigation revealed segmentation differences in the basal slices, which had a significant impact on the volumetric assessment.



	Clinical Result	Gold	Reader2	)iff(Gold, Reader2
0	LVESV	65.0	61.0	4.0
1	LVEDV	153.1	153.2	-0.1
2	RVESV	73.8	74.2	-0.4
3	RVEDV	159.7	162.1	-2.4
4	LVSV	88.1	92.2	-4.1
5	LVEF	57.4	60.1	-2.7
6	RVSV	85.9	87.9	-2.0
7	RVEF	54.0	54.8	-0.8

rv_endo

	slice	SAX RVED ml diff	SAX RVES ml diff	\X RVED abs ml d	\X RVES abs ml d	SAX RVED DSC	SAX RVES DSC	SAX RVED HD	SAX RVES HD	‹ RVED has_contc	‹
0	0	0.0	0.0	0.0	0.0	100	100	0.0	0.0	False	F
1	1	0.0	0.0	0.0	0.0	100	100	0.0	0.0	False	F
2	2	-13.8	0.0	13.8	0.0	51	100	28.0	0.0	True	F
3	3	-0.2	0.0	0.2	0.0	96	100	1.4	0.0	True	F
4	4	0.3	3.3	0.3	3.3	96	62	1.7	16.5	True	T

Figure 4: Tracing Difference from Statistical Plots to Annotation Differences

Attribution: adapted from "Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging" by Hadler et al. 2022, https://www.nature.com/articles/s41598-022-10464-w, Licensed under a Creative Commons Attribution 4.0 License.

Caption: Two Lazy Luna tabs are shown on the left. The top tab focuses on clinical parameter statistics: top left shows clinical parameter averages for each reader and their differences in a table. Top right shows a paired boxplot for the selected clinical parameter, first reader on top, second below. Bottom left show a QQ-plot. The bottom right RVEDV Bland Altman plot is magnified on the right to show differences as assessed by both readers. Case x was selected by the user to open the lower tab, which presents the outlier's segmentations. The top part of the second tab shows metric values for the segmentation comparisons, the lower part a segmentation comparison, which is magnified on the right. The first reader's segmentations are in red (first subplot); the second's in blue (third subplot), and their agreement is displayed in the second subplot with green referencing area overlap between both readers.

Legend: RVEDV: right ventricular end-diastolic volume, QQ: quantile-quantile

Two tables were automatically generated by LL during the multilevel reader comparison. The first shows all SAX cine clinical parameter value averages and standard deviations, as well as segmentation metrics that affected the clinical parameters (Table 2). The second shows where segmentation difficulties statistically accumulate in cardiac geometry, subdivided by contour type and basal, midventricular or apical slices (Table 3). As illustrated qualitatively in Figure 4, RV endocardial contours in basal slices were shown to be more difficult to contour than other regions and have higher average volumetric impacts (3.1ml per slice, Table 3).

Table 2: **Reader Comparison: Clinical Parameters and Segmentation Metrics**

Attribution: adapted from "Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging" by Hadler et al. 2022, https://www.nature.com/articles/s41598-022-10464-w, Licensed under a Creative Commons Attribution 4.0 License.

Clinical Parameter Difference / Metric	**Mean**	**Std**
LVEF [%]	-2.7	2.9
LVEDV [ml]	-0.1	2.7
LVESV [ml]	4	4.4
Dice [%]	94	3
HD [mm]	0.7	0.3
LVM [g]	-1	4
Dice [%]	91	7
HD [mm]	0.8	0.5
RVEF [%]	-0.8	3.1
RVEDV [ml]	-2.4	11.1
RVESV [ml]	-0.4	6.2
Dice [%]	90	5
HD [mm]	1.6	0.7
All Contour Types		
Dice [%]	92	4
HD [mm]	1.1	0.5
Legend: LV: left ventricle, RV: right ventricle, LVM: Left ventricular myocardial mass, EF: ejection fraction, EDV: end-diastolic volume, ESV: end-systolic volume, Dice: Dice similarity coefficient, HD: Hausdorff metric, Std: Standard deviation		

Table 3: **Reader Comparison: Clinical Parameters and Segmentation Metrics by Cardiac Location**

Attribution: adapted from "Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging" by Hadler et al. 2022, https://www.nature.com/articles/s41598-022-10464-w, Licensed under a Creative Commons Attribution 4.0 License.

Position	Metric	LV Endocardium	LV Myocardium	RV Endocardium
Basal	Dice [%]	88	87	72
	HD [mm]	1.9	2.1	8.1
	Abs. ml diff [ml]	1.4	0.9	3.1
Midv	Dice [%]	97	91	94
	HD [mm]	0.8	1	2
	Abs. ml diff [ml]	0.3	0.4	0.6
Apical	Dice [%]	84	74	83
	HD [mm]	0.2	0.5	0.2
	Abs. ml diff [ml]	0.2	0.5	0.2

Legend: LV: left ventricle, RV: right ventricle, Midv: midventricular, EF: ejection fraction, EDV: end-diastolic volume, ESV: end-systolic volume, Dice: Dice similarity coefficient, HD: Hausdorff metric, Abs. ml diff: absolute millilitre difference per slice, Std: Standard deviation

3.2 Software Architecture Design

3.2.1 Accessibility and Product Independence

LL exclusively builds on open-source libraries as prescribed in Methods. LL itself is offered as open-source software on Github (https://github.com/thadler/LazyLuna), and was tested on 64-Bit systems of macOS 10.14.6 and 12.6.8, Windows 10 Home and Ubuntu 20.04 and worked for Python versions 3.8 and above.

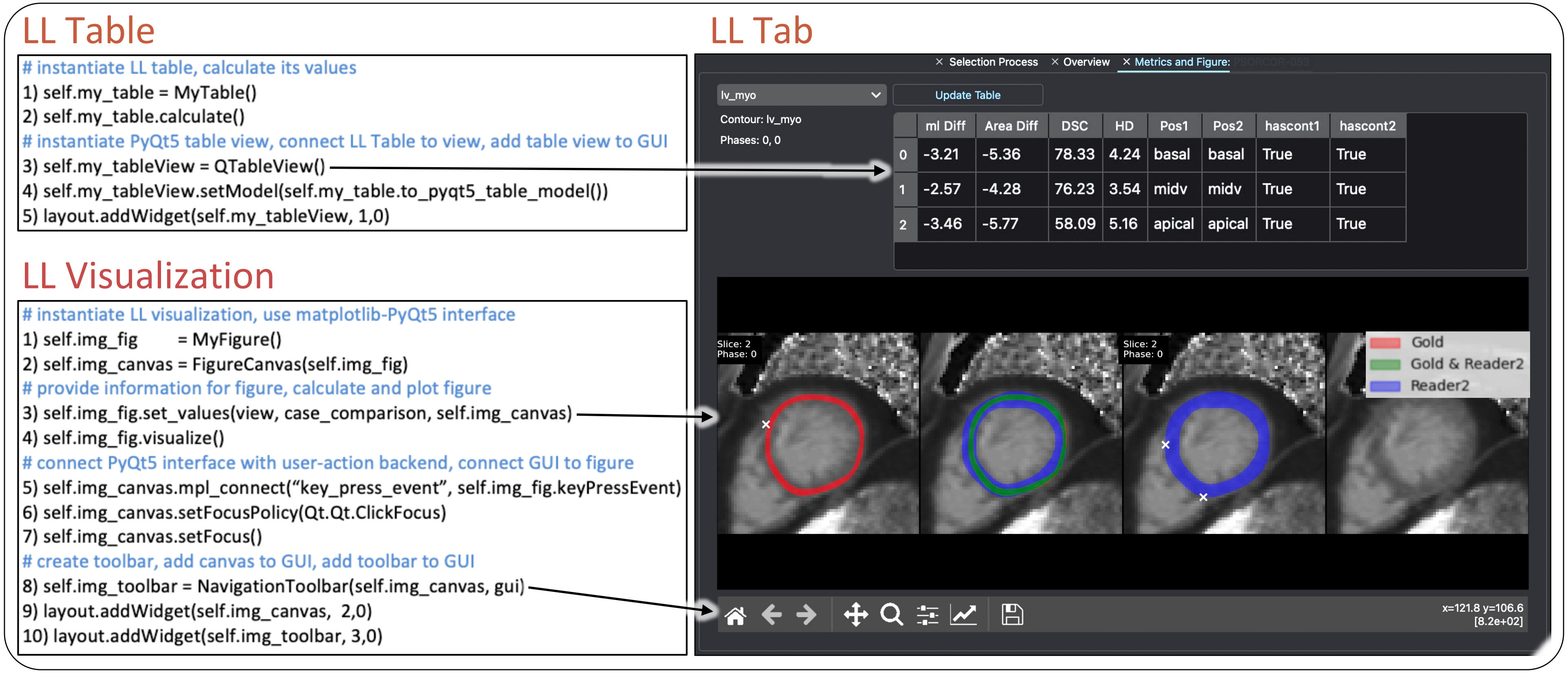
3.2.2 Usability and Target Groups

The two extensions below were developed in an agile development procedure with feedback from clinicians on the utility of extensions.

3.2.3 Extendibility

Plug-in Scheme

Tables and visualizations can be added to tabs following a plug-in scheme (Figure 5).



LL Table

```
# instantiate LL table, calculate its values
1) self.my_table = MyTable()
2) self.my_table.calculate()
# instantiate PyQt5 table view, connect LL Table to view, add table view to GUI
3) self.my_tableView = QTableView()
4) self.my_tableView.setModel(self.my_table.to_pyqt5_table_model())
5) layout.addWidget(self.my_tableView, 1,0)
```

LL Visualization

```
# instantiate LL visualization, use matplotlib-PyQt5 interface
1) self.img_fig     = MyFigure()
2) self.img_canvas = FigureCanvas(self.img_fig)
# provide information for figure, calculate and plot figure
3) self.img_fig.set_values(view, case_comparison, self.img_canvas)
4) self.img_fig.visualize()
# connect PyQt5 interface with user-action backend, connect GUI to figure
5) self.img_canvas.mpl_connect("key_press_event", self.img_fig.keyPressEvent)
6) self.img_canvas.setFocusPolicy(Qt.Qt.ClickFocus)
7) self.img_canvas.setFocus()
# create toolbar, add canvas to GUI, add toolbar to GUI
8) self.img_toolbar = NavigationToolbar(self.img_canvas, gui)
9) layout.addWidget(self.img_canvas,  2,0)
10) layout.addWidget(self.img_toolbar, 3,0)
```

LL Tab

Contour: lv_myo
Phases: 0, 0

	ml Diff	Area Diff	DSC	HD	Pos1	Pos2	hascont1	hascont2
0	-3.21	-5.36	78.33	4.24	basal	basal	True	True
1	-2.57	-4.28	76.23	3.54	midv	midv	True	True
2	-3.46	-5.77	58.09	5.16	apical	apical	True	True

Figure 5: Lazy Luna Plug-in Scheme

Attribution: adapted from "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.

Caption: The left presents LL table and visualization code examples, the right shows the resulting GUI tab. For the Table: the first code line instantiates the LL table, the second calculates its cell values, the third instantiates the GUI table, line four connects the LL table to GUI table, line five makes the GUI table visible. For the visualization: the visualization is instantiated in code line one, the interface canvas in line two, interface parameters are set in line three, the figure is calculated in line four, user-interactions are connected in lines five to seven, the toolbar is added to the GUI in line eight, and then added to the GUI in lines nine and ten. On the right, the GUI is shown with the table on top and the figure below. Legend: LL: Lazy Luna, GUI: graphical user interface

3.2.4 Extension to T1 Parametric Mapping

In development cycles with clinicians parametric T1 mapping requirements were defined, including the GLOBAL_T1 value and the American Heart Association (AHA) model [65]. The GLOBAL_T1 value is the average of all pixel values within the myocardial contours across all slices, while the AHA model calculates T1 averages of pixels within 16 myocardial segments. This required extending the Annotation class with functions capable of identifying pixels within the myocardial contour, which was performed with Rasterio's rasterize function [66].

Calculating the AHA model and an AHA differences model required extensions to the Annotation and the Category class (Figure 6). The Annotation class was extended to allow for calculating AHA segments by using the insertion point and the LV myocardial segmentation. Within these extensions, the LV endocardial centroid and the insertion point are used to divide the LV myocardial segmentation into several segments (6 for basal and mid ventricular, 4 for apical slices). Then mean values are calculated for pixels within the individual segments.

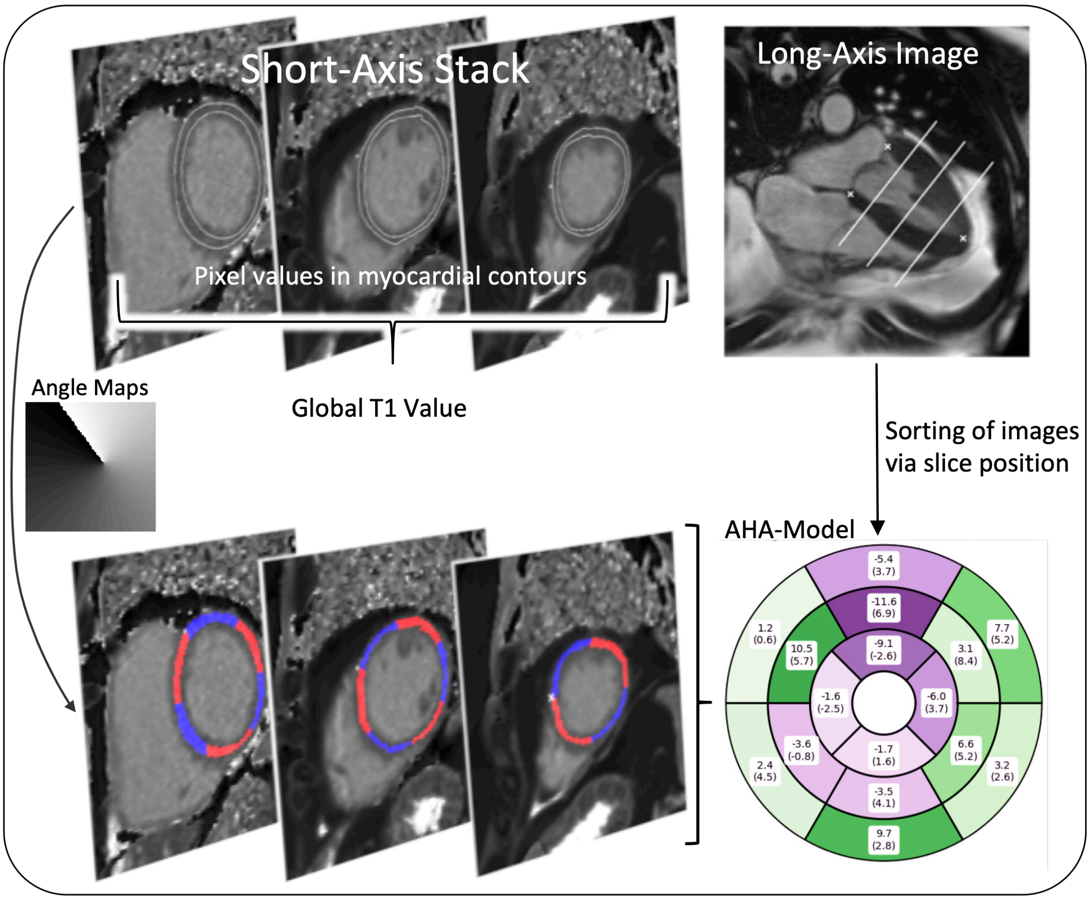


Figure 6: Lazy Luna Calculation of the AHA model
Attribution: adapted from "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.
Caption: Short-axis parametric T1 mapping images with LV myocardial delineations and insertion points on the top left. Bottom left, myocardial segment masks are calculated (red and blue) from the image annotations. On the top right, the images and segmentation masks are assigned to basal, midventricular or apical locations, depending on their spatial relationship to the extent and apical points in a long-axis view of the heart. The bottom right shows the AHA model by assigning the segments into their respective bins and calculating the average. The rings correspond to basal, midventricular and apical positions (outside to inside).
Legend: LV: left ventricle, AHA: American heart association

Building on these code adaptions two new LL Visualisation classes were implemented and added to LL to visualise the AHA models for both cases, and an AHA difference model to illustrate differences that emerge due to different annotation choices by both readers (Figure 7).

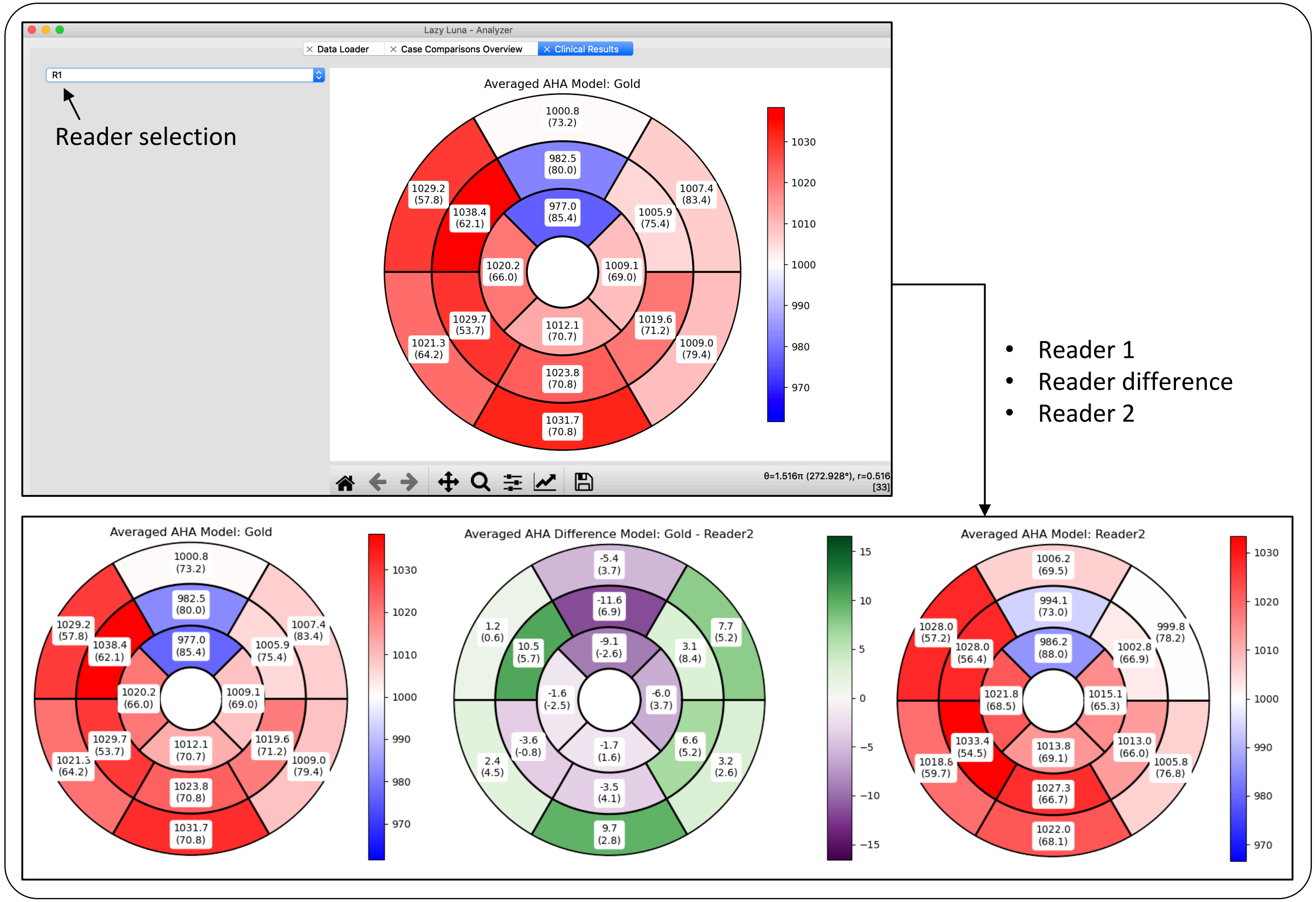


Figure 7: AHA Model

Attribution: adapted from "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.

Caption: On the top, a Lazy Luna tab with a reader's AHA model average for several patients is shown. The reader was selected on the upper left of the GUI. Below three figures generated from this tab are presented (from left to right): First, the average AHA model for all cases for the first reader is presented. Second, the average of differences between the first and second reader is visualized. Third, the average AHA model for all cases for the second reader is shown.

Legend: AHA: American heart association, GUI: graphical user interface

3.2.5 Extension to Late Gadolinium Enhancement

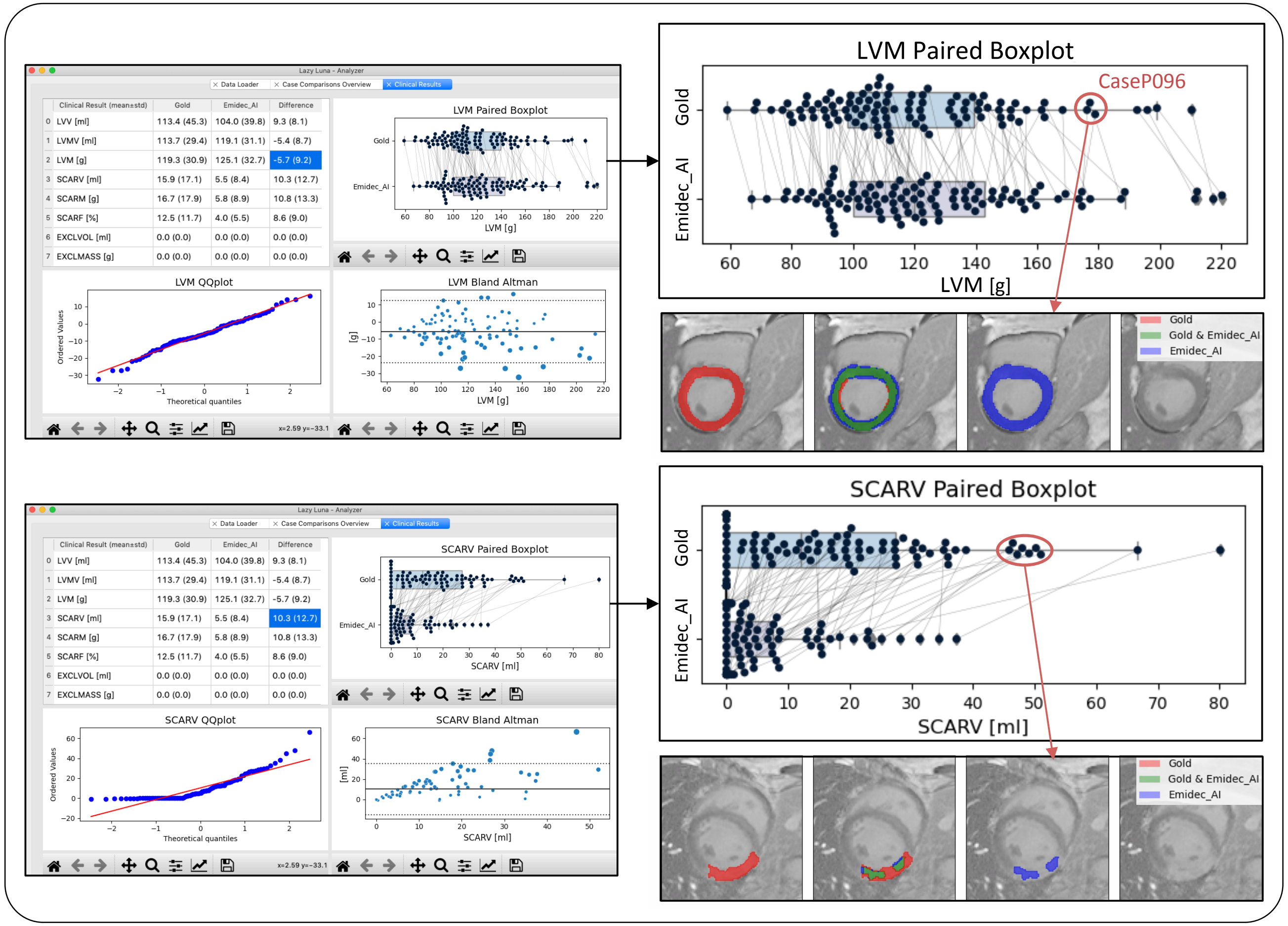
The following subchapter is a revised version of the Results chapter from the publication "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging" by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.

The annotation class required no changes to generalize to LGE imaging. New contour types (scar, excluded area, no reflow) were integrated into the annotation class, which allowed for dealing with scars, exclusions and no reflow areas, which permitted the implementations of LGE clinical parameters, including the scar volume, mass and fraction, excluded volume and mass as well as no re-flow volume and mass. "LL also required a new View class, the SAX_LGE_View to refocus the cases on the images and clinical parameters" [53].

Extension to the Emidec dataset

The Emidec dataset comprises LGE images and segmentation masks stored in Nifti format. Nifti provides image meta information and the image. Meta information "contains voxel width, height, and depth, image sizes, and the number of phases and slices" [53]. This information was converted from Nifti format to DICOM files. "The annotation files were generated from the Nifti file masks" [53]. Nifti voxel segmentations were outlined as Shapely polygons, converted to LL annotation format and stored as pickle files, assigning Shapely geometries to contour names. The code was published as a Jupyter Notebook (interactive programming environment) in the repository.

An under-trained UNet (called Emidec_AI) predicted segmentation masks for myocardium, scar and no reflow tissue. The Emidec_AI predicted the training set's segmentations. LL was used to compare the readers in Figure 8.



	Clinical Result (mean±std)	Gold	Emidec_AI	Difference
0	LVV [ml]	113.4 (45.3)	104.0 (39.8)	9.3 (8.1)
1	LVMV [ml]	113.7 (29.4)	119.1 (31.1)	-5.4 (8.7)
2	LVM [g]	119.3 (30.9)	125.1 (32.7)	-5.7 (9.2)
3	SCARV [ml]	15.9 (17.1)	5.5 (8.4)	10.3 (12.7)
4	SCARM [g]	16.7 (17.9)	5.8 (8.9)	10.8 (13.3)
5	SCARF [%]	12.5 (11.7)	4.0 (5.5)	8.6 (9.0)
6	EXCLVOL [ml]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
7	EXCLMASS [g]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)

	Clinical Result (mean±std)	Gold	Emidec_AI	Difference
0	LVV [ml]	113.4 (45.3)	104.0 (39.8)	9.3 (8.1)
1	LVMV [ml]	113.7 (29.4)	119.1 (31.1)	-5.4 (8.7)
2	LVM [g]	119.3 (30.9)	125.1 (32.7)	-5.7 (9.2)
3	SCARV [ml]	15.9 (17.1)	5.5 (8.4)	10.3 (12.7)
4	SCARM [g]	16.7 (17.9)	5.8 (8.9)	10.8 (13.3)
5	SCARF [%]	12.5 (11.7)	4.0 (5.5)	8.6 (9.0)
6	EXCLVOL [ml]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
7	EXCLMASS [g]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)

Figure 8: **Emidec Artificial Intelligence Investigation**

Attribution: adapted from “Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging” by Hadler et al. 2023, https://www.sciencedirect.com/science/article/pii/S0169260723002808, Licensed under a Creative Commons Attribution 4.0 License.

Caption: The Clinical Results tab (top left) shows clinical parameter averages of the gold standard, the Emidec_AI, and their differences. The paired boxplot is enlarged on the upper right. The two readers (Gold top, Emidec_AI bottom) are presented as boxplots with the cases‘ LVM values plotted as dots. Lines connect the case dots to visualize case-specific reader differences. Below, the reader contours of CaseP096 are presented, showing the overestimation of the epicardial contour, explaining LVM differences. The bottom tab’s sub-figures provide an analogous analysis for the Emidec_AI‘s SCARV estimation. The paired boxplot reveals that the Emidec_AI underestimates scar volumes. The lower plot shows the Emidec_AI has not learned to estimate the full scar.

Legend: LVM: left ventricular myocardial mass, SCARV: scar volume

4 Discussion

4.1 Short Summary of Results

The “Software Prototype Development” and the “Software Architecture Design” publications showed that a flexible and extendible semi-automated multilevel reader comparison software tool could be developed for CMR quality control and published as open-source software. The software Lazy Luna was designed as a software package to solve a number of intermingling calculations (i.e. metrics, clinical parameters, statistics), visualisations, relationships between analysis levels (i.e. reader biases on clinical parameters and annotation choices) and tabs of the user interface (e.g. connecting tabs with statistical visualisations to tabs illustrating a reader comparison on an individual case). The main result of the “Software Prototype Development” publication is demonstrating the feasibility of implementing such a comparison tool. On 13 short-axis cine cases the software calculated segmentation metrics, clinical parameters and reader biases. LL performed statistical procedures, and offered visualisations of annotation differences and statistics. However, prototyping also revealed insufficiencies with the underlying software design, such as its exclusive dedication to SAX cine imaging. This was addressed in the follow-up publication. The main result of publishing the “Software Architecture Design” was a generic software backend and frontend for simple extensions to new imaging techniques by streamlining the implementation of new figures, tables and tabs for the graphical user interface. This was demonstrated by extending LL to T1 mapping and late-gadolinium enhancement imaging with visualisations, tables and tracing methods, as well as testing the software on an interobserver T1 dataset, and a publicly available LGE dataset with scar contours where expert annotations were compared to a convolutional neural network.

Some studies that applied LL are presented in the discussion. These are not the focus of this thesis, but demonstrate the software’s utility as a quality assurance support tool for CMR. In the following subchapters we show how LL helped reveal connections between confounders, annotation and clinical parameter differences (4.2). Then we focus on the relevance of high-quality data (4.3) as LL showed that clean data is key to quality assurance in CMR. Following the Limitations (4.4), the outlook is presented (4.5), encompassing: knowledge extraction in data storage systems (4.5.1), the clinical integration of AIs

(4.5.2), establishing new clinical parameters (4.5.3) and training newcomers as they enter clinical routine (4.5.4).

4.2 Annotation Difficulties, Confounders and Future Standardisation

Developing LL was accompanied by numerous software applications in various scenarios. However, regardless of the application, LL consistently revealed substantial annotation differences between readers. Reasons for annotation differences are manifold and originate along the entire CMR imaging chain. Some differences suggest contradictory beliefs about the cardiac structures visible in the images [48]. For example, in the “Software Prototype Development” publication SAX cine basal slice decisions occasionally caused large volumetric differences between readers. These were due to partial volume effects, which genuinely present mutually exclusive cardiac structures in the same image (e.g. ventricle and atria). During the “Software Architecture Design” the integration of myocardial border voxels was revealed as difficult. In order to determine precise myocardial parametric mapping values, contours must be rasterised (generating pixel masks from polygons) and determined as either belonging to the myocardium or not. Beyond being a non-trivial implementation and software design problem, human readers face this problem every time they segment parametric mapping images, choosing more or less conservative myocardial segmentations to avoid voxels containing fat and blood. Occasionally, lacking consistency and missing annotations of cardiac structures can be explained by them being overlooked such as human readers segmenting the papillary muscles in slices n-1 and n+1 but not in-between, in slice n. Of course, for AIs this explanation does not hold. In the “Software Architecture Design” publication, an AI was trained to segment the LV myocardium as well as scar and no reflow tissue. The AI performed well on myocardium segmentation, but poorer on scar tissue detection, and did not detect any no reflow regions. This roughly reflected the prevalence of the different classes, but was (as intended in the study) due to an insufficient training of the network. In 2023 Ammann et al. applied LL to analyse commonalities and differences between three popular CNN architectures [48]. All architectures were trained in a comparable environment to segment SAX cine images, and evaluated on 29 test cases. Two CNN architectures were similar for clinical parameters across the board, typically within predefined tolerance ranges; one had a poorer performance, and often exceeded tolerance ranges. However, for all CNN architectures, basal slice volume differences were the foremost origin of ventricular volume differences, however, different from human readers, CNNs were reluctant to decide for or against slice segmentation, instead segmenting basal and apical slices partially and implausibly.

The point of listing these annotation difficulties is not to instantaneously solve these issues but to show that qualitative investigations of annotation failures are important because in order to address them appropriately they must be pinpointed in their causal origin. Depending on the origin of a segmentation difficulty, adjusting different steps in the CMR imaging chain seems advantageous to their resolution. On the one hand, partial volume artefacts occur with large slice thicknesses when different structures or tissues are encompassed within the same voxel and produce intermediate values. For example, when mutually exclusive cardiac structures are represented in the same SAX cine images, no amount of post-processing guidelines will resolve the discrepancies. Rather, this segmentation difficulty can be better addressed by increasing resolution along the z-axis of the stacks, such as would be offered by 3D sequences [67]. This would decrease the voxel depth of individual slices, which leads to smaller volumetric impacts of individual slices and simultaneously to fewer cardiac structures being averaged into the same image. On the other hand, some annotation differences may point towards the need for more annotation standardisation, such as differences in papillary muscle inclusion choices, in basal slice choices by expert readers or CNNs violating cardiac geometry in basal slices, which may result from training on inconsistent datasets. Artefacts in images must similarly be differentiated by their cause [16]. Breathing artefacts make image annotation difficult and may be caused by severely ill patients or healthy volunteers who struggle with breathing commands. These artefacts can be addressed either by using faster sequences that require fewer breath holds, such as compressed sensing sequences or by developing novel reconstruction methods that attempt to correct for K-space data affected by motion artefacts [15,23,68,69].

Several of these annotation difficulties may require flagging as they unveil themselves during the CMR examination and post-processing. Breath hold problems can be tackled with appropriate sequences; artefacts can be detected during reconstruction with signal-to-noise ratio calculations or with classification CNNs after reconstruction, and programs that assess whether annotations respect cardiac geometry constraints can identify implausible annotations. Such consistency checks and red flagging of potential issues along the CMR imaging chain can provide improvement suggestions and increase segmentation plausibility while catching post-processing errors during stressful clinical routine.

4.3 High-Quality Data and CMR Quality Assurance

4.3.1 Heterogeneous Data

LL supports CMR Quality Assurance (QA) tasks by comparing a reader or an AI to an expert reader. However, in order to attain good QA assessments, excellent and representative data is necessary. In the last decade CMR has exploited the heterogeneity of data in order to address confounders, such as patient characteristics, including age, sex, ethnicity and diseases, over scanning characteristics, such as vendors, field strengths, coils, sequence implementations and image reconstruction algorithms to post-processing steps, including different sites, post-processing software, and readers. For clinicians, disease heterogeneity is axiomatic to disease differentiation, diagnosis and establishing reference values [70,71]. To clinical researchers heterogeneity is relevant because the communicability of results and diagnostic approaches builds on the comparability of normal values between sites. As confounders interfere with the comparability of site-specific normal values, Z-score normalization and multi-site evaluations attempt to compensate or evaluate the communicability of parameters and their normal values [20,21,40].

Since the introduction of CNNs, dataset sizes have greatly increased in CMR. Next to dataset size, data heterogeneity has proven key to improving AI performance in real world scenarios, which require models that are hardened against inevitable confounders and accompanying domain shifts [32,33,72]. As the first CMR segmentation competition datasets were produced before the arrival of CNNs, they focussed on individual cardiac structures in SAX cine, such as LV and RV segmentation datasets in 2009-2012 - at the time difficult computer vision problems [73,74]. However, as CNNs overtook traditional methods in semantic segmentation tasks, datasets increased in difficulty. In 2017, the ACDC dataset offered segmentations of both ventricles on healthy and pathological hearts [32]. In 2020 and 2022, the M&Ms dataset boosted the trend towards mirroring clinical reality by increasing dataset heterogeneity with multiple centres, vendors and diseases in their segmentation competition [33].

Although the engineering of datasets towards real-world data reflects good machine learning practice [75], the gap between segmentation competitions and clinical routine data is far from closed – a recent publication on confounder impact in clinical routine pointed towards unknown confounders having long-term influences on clinical parameter stability [76]. Datasets should reflect the full range of available imaging techniques with clinical utility, including SAX and LAX cine, parametric mapping techniques, LGE and 4D flow. Datasets should also orient themselves towards the diverse set of patients encountered in clinical routine. This includes diverse representations of diseases, but should also include patients covering several age groups, evenly distributed between both genders and ethnicities, as well as integrating different scanners, sites and post-processing software. The patient-focused heterogeneity is necessary because the statistical averages of clinical routine patients vary wildly across the world, which means that extracting cohorts from one clinic will inevitably integrate spurious correlations reflecting the region's bias in diseases, ages, genders, ethnicity, payment plans of the health case system and diagnostic procedures in practice. Future datasets should include these confounders in order to best mirror the diversity of clinical reality.

4.3.2 Clean Data

Provided a heterogeneous dataset that closely represents real-world data, expert annotations reveal themselves as essential to CMR QA. Reproducible and accurate annotations lead to less variance and bias in the dataset, which in turn leads to training more proficient AIs and contributes to straightforward evaluations of trainees on dedicated datasets. Annotation variability remains significant in CMR datasets [38], and makes training and benchmarking AIs difficult [77]. Recently, segmentation competitions, such as ACDC and M&Ms, have aimed at generating excellent annotations with several expert readers agreeing on annotations, thus defining a gold standard dataset [32,33]. Nonetheless, these competitions typically store their segmentations as pixel masks and not as polygons, thus losing sub-pixel resolution that was integrated in the post-processing software - this may be particularly relevant for thin myocardia, which can be as thin as individual pixels. At the same time, reader agreement may be beneficial to reducing overlooked contours and other careless mistakes, but seems less promising at eliminating real disagreements that result from unclear data, that different readers genuinely understand differently [38]. Annotation variance affects the training of AIs, which learn reader indecisiveness as fragmented annotations “in-between” two plausible annotations of the same image [48,50]. This reflects their tendency to reduce the bias between their outputs and the image’s multiple plausible annotations. However, although the two individual annotations may represent plausible solutions, their average may violate cardiac geometry constraints. As data in CMR grows exponentially, the annotation differences themselves become a valuable resource for QA tasks, such as developing annotation repair methods and reproducibility assessments.

4.4 Limitations

LL requires user intervention to recognise imaging techniques (i.e. SAX/LAX cine or parametric mapping images). This is performed as a semi-automated procedure for each case in order to ensure correct image classification. Currently, LL is limited to comparing two readers to each other. Future versions of LL should allow the comparison of multiple readers.

4.5 Outlook

4.5.1 Data Storage and Knowledge Extraction

The growing amount and complexity of CMR data requires more expressive storage options to do them justice. Picture archiving and communications systems (PACS) and the DICOM standard were optimised for clinical needs and patient-oriented requests, such as searching for acquisitions by date. As CMR moves through cycles of research investigations and clinical deployments, a multitude of confounders are investigated for effects on annotations and clinical parameters, with AI deployments offering annotations of all phases and slices in cine imaging, potentially establishing new clinical parameters and humanly unverifiable numbers of annotations. These research endeavours are unlocking a wealth of data unsuited to traditional PACS. First, PACS limit the kinds of data that can be stored together, and second, they limit the options to query this data efficiently and extract knowledge [78–80]. Knowledge extraction from databases could be used for research purposes, such as testing and generating hypotheses from vast stores of data (i.e. data mining). QA tools like LL could be redesigned to interface with such a database. In order to fully harness the available data, better accommodate its integration with data from other modalities, and offer explorative data analysis, the underlying databases should be engineered towards the emerging research environment.

4.5.2 Clinical Integration of Artificial Intelligence

Another data science task that should be adjusted to the QA concerns of medical imaging is the training and evaluation of AI algorithms. Currently, AI training is exclusively being treated as a Deep Learning problem, and although this is an important aspect, it is inherently incomplete given the complexity of the domain in which the AIs are intended to function. These tasks could also be performed on a dedicated database that integrates an ever-growing number of confounders, so that the AI can expand its robustness to confounders as CMR itself expands. And while AIs calculate impressive clinical parameters, in the range of expert reader deviations, they continue to produce nonsensical annotations. These difficulties cause distrust and hamper their integration into clinical routine. Recent FDA proposals suggest that clinical AI integration may be best served with a constant monitoring-approach [75,81]. In the following paragraphs we will, first, outline steps for the post-processing with AIs, second, deal with AI-output acceptability, and third, illustrate a monitoring environment for AIs in clinical routine.

Images may contain artefacts that make their interpretation misleading. Artefacts should be automatically identified and reported. Following this, irrelevant images are excluded from the segmentation task (e.g. SAX cine images outside the ventricles). These image classification tasks can be tackled with CNN image classifiers [82,83]. The relevant cardiac structures should be extracted from the image with bounding box detection [76,84]. Following these two preprocessing steps, the images are segmented with segmentation CNNs, such as U-Nets [45,85–87]. Following the segmentation procedure, output masks are post-processed. As the last step of the segmentation CNN is typically a sigmoidal function, which maps each mask pixel value to a floating point number between 0 and 1, the output is thresholded to label pixels as either belonging to a cardiac structure or not. This thresholding can lead to “stray pixels” or even stray structures [50]. In order to exclude these stray structures, the largest connected component is often selected, setting other mask pixels to zero. However, the necessity of the largest connected component to exclude stray pixels/structures implies that an understanding cardiac geometry is not inherent to the CNN itself.

Even the largest connected component can reflect an indecisive segmentation decision, leading to the fragmented segmentations that violate cardiac geometry and must be dealt with separately. Several approaches integrate cardiac geometry constraints into the AIs, such as shape-priors for loss functions [88,89] during training, shape-constraints on viable geometries [90,91], or by replacing the CNN segmentation with similar expert segmentations from other hearts (similarity determined on an embedding of predicted and expert segmentations) [92]. In this case, segmentation error estimation may help users identify segmentations with difficulties. Methods like Monte-Carlo Dropout segmentation generation [93], reverse classification accuracy[94], linear combinations of ensemble Dice estimations [95] and Bayesian networks [49,50] intend to offer the user segmentation quality estimation. Explainability methods like GradCam [96] have been tested successfully on brain MRI scans, and may provide the user with an insight into which CMR image regions are confusing. Future methods may evaluate the 3D plausibility of segmentation stacks.

A clinical AI should be embedded in a monitoring and self-updating environment with a QA database. Data from clinical routine must be accessible for AI training and evaluation to ensure good performance and robustness towards confounders [81]. After deploying the clinical AI, it performs on clinical routine data, while simultaneously being monitored so that failed annotations are caught and corrected. These corrected annotations are added to the QA database. By storing training and QA cases in the database (and expanding them with clinical cases over time), the AI can be continuously validated on the database to check for consistent or improving performance. Such an expansion of training cases will further allow for continuous updates of a clinical AI to reflect annotation guideline evolution, newly diagnosable and differentiable diseases, without risking undetected model performance deterioration.

4.5.3 Establishing and Assessing Novel Clinical Parameters

As CMR evolves to exploit advantages of AI segmentation, new clinical parameters that require vast numbers of segmentation are established. Such innovative parameters must be investigated for their reproducibility and plausibility – a task that LL supports. For example, in 2023 Gröschel et al. used LL to evaluate the reliability of AI based quantification of myocardial strain [97]. Myocardial strain was calculated in SAX and LAX cine views by generating LV myocardial contours and extent points for all phases in all views and “tracking” the displacement of myocardial features to compute strain values. To this end, LL was used to assess contour plausibility by comparing AI segmentations to expert contours quantitatively, and calculate statistical differences between AI and expert strain assessments. By differentiating the influences of contouring proficiency from feature tracking, LL allowed to isolate the effect of the feature-tracking algorithm. The future of CMR is bound to include new clinical parameters that require high-dimensional representations extracted from more images than could be manually segmented efficiently. For example, blood flow curves (for LV, RV, and atrial chambers) are calculated by segmenting all cine images of SAX and LAX views, and were effectively applied by Bello et al. in 2019 to predict mortality rates more accurately than conventional methods could [98]. The same holds true for CMR shifting from slice-based imaging to 3D volumetric imaging. 3D imaging techniques provide isotropic high-resolution images that capture intricate cardiac structures, but they will also require hundreds of image segmentations to delineate cardiac structures. Extensions of LL to 3D sequences will be a future undertaking to allow for feedback loops as AIs and new parameters are assessed for their clinical utility.

4.5.4 Training CMR Newcomers

Within the working group LL is used to provide trainee feedback with semi-automatically generated reports, which often show the steep learning curve necessary to master CMR post-processing. While LL is capable of identifying clinical parameter deviations between trainee and expert, as well as exposing segmentation differences that caused the parameter deviations, it also shows that LL requires user-intervention to describe annotation differences for the trainee. On the one hand, learning CMR encompasses far more than annotating images, it also includes identifying and describing structures, patterns, atypical morphologies, and recognising pathologies in images. On the other hand, the training procedure is far from replaced; LL remains a semi-automated tool, leaving the communication of mistakes to the supervisor or LL-user. Future research will model the training of CMR newcomers as a “gamified” learning experience, in which the trainee is presented with a base set of cases to annotate. As the trainee progresses through the cases, annotation proficiency and clinical parameter deviations are tracked. After the trainee has annotated several cases, the trainee receives an automatic report. Certain segmentation failures could be automatically classified through shape matching and linked to explanation/training videos on segmentation choices by experts before the training continues.

5 Conclusions

The developed multilevel reader comparison software, Lazy Luna, was successfully prototyped for short-axis cine imaging, then thoroughly designed and implemented to be extendible to new imaging techniques and generalize over typical quality assurance tasks in Cardiovascular Magnetic Resonance. The presented studies show that quality assurance tool design and implementation is both possible and feasible. The numerous applications presented in results and the discussion revealed deeper insights into why standardisation in CMR is challenging and which paths are most promising. Future research will focus on generalising the reader comparison software to multiple other imaging techniques and confounders, in order to support and enable strong AI research and training improvements in CMR.

Reference List

1. Ridgway JP. Cardiovascular magnetic resonance physics for clinicians: part I. J Cardiovasc Magn Reson. 2010 Dec;12(1):71.

2. Biglands JD, Radjenovic A, Ridgway JP. Cardiovascular magnetic resonance physics for clinicians: part II. J Cardiovasc Magn Reson. 2012 Dec;14(1):66.

3. 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 : Society for Cardiovascular Magnetic Resonance (SCMR): Board of Trustees Task Force on Standardized Post-Processing. J Cardiovasc Magn Reson. 2020 Mar 12;22(1):19.

4. on behalf of SCMR Clinical Trial Writing Group, Puntmann VO, Valbuena S, Hinojar R, Petersen SE, Greenwood JP, Kramer CM, Kwong RY, McCann GP, Berry C, Nagel E. Society for Cardiovascular Magnetic Resonance (SCMR) expert consensus for CMR imaging endpoints in clinical research: part I - analytical validation and clinical qualification. J Cardiovasc Magn Reson. 2018 Dec;20(1):67.

5. Von Knobelsdorff-Brenkenhoff F, Schulz-Menger J. Cardiovascular magnetic resonance in the guidelines of the European Society of Cardiology: a comprehensive summary and update. J Cardiovasc Magn Reson. 2023 Jul 24;25(1):42.

6. Zeppenfeld K, Tfelt-Hansen J, De Riva M, Winkel BG, Behr ER, Blom NA, Charron P, Corrado D, Dagres N, De Chillou C, Eckardt L, Friede T, Haugaa KH, Hocini M, Lambiase PD, Marijon E, Merino JL, Peichl P, Priori SG, Reichlin T, Schulz-Menger J, Sticherling C, Tzeis S, Verstrael A, Volterrani M, ESC Scientific Document Group, Cikes M, Kirchhof P, Abdelhamid M, Aboyans V, Arbelo E, Arribas F, Asteggiano R, Basso C, Bauer A, Bertaglia E, Biering-Sørensen T, Blomström-Lundqvist C, Borger MA, Čelutkienė J, Cosyns B, Falk V, Fauchier L, Gorenek B, Halvorsen S, Hatala R, Heidbuchel H, Kaab S, Konradi A, Koskinas KC, Kotecha D, Landmesser U, Lewis BS, Linhart A, Løchen ML, Lund LH, Metzner A, Mindham R, Nielsen JC, Norekvål TM, Patten M, Prescott E, Rakisheva A, Remme CA, Roca-Luque I, Sarkozy A, Scherr D, Sitges M, Touyz RM, Van Mieghem N, Velagic V, Viskin S, Volders PGA, Kichou B, Martirosyan M, Scherr D, Aliyev F, Willems R, Naser N, Shalganov T, Milicic D, Christophides T, Kautzner J, Hansen J, Allam L, Kampus P, Junttila J, Leclercq C, Etsadashvili K, Steven D, Gatzoulis K, Gellér L, Arnar DO, Galvin J, Haim M, Pappone C, Elezi S, Kerimkulova A, Kalejs O, Rabah A, Puodziukynas A, Dimmer C, Sammut MA, David L, Boskovic A, Moustaghfir A, Maass AH, Poposka L, Mjolstad OC, Mitkowski P, Parreira L, Cozma D, Golukhova E, Bini R, Stojkovic S, Hlivak P, Pernat A, Castellano NP, Platonov PG, Duru F, Saadi ARA, Ouali S, Demircan S, Sychov O, Slade A. 2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. European Heart Journal. 2022 Oct 21;43(40):3997–4126.

7. Leiner T, Bogaert J, Friedrich MG, Mohiaddin R, Muthurangu V, Myerson S, Powell AJ, Raman SV, Pennell DJ. SCMR Position Paper (2020) on clinical indications for cardiovascular magnetic resonance. J Cardiovasc Magn Reson. 2020 Dec;22(1):76.

8. Messroghli DR, Moon JC, Ferreira VM, Grosse-Wortmann L, He T, Kellman P, Mascherbauer J, Nezafat R, Salerno M, Schelbert EB, Taylor AJ, Thompson R, Ugander M, van Heeswijk RB, Friedrich MG. Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: A consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI). J Cardiovasc Magn Reson. 2017 Dec;19(1):75.

9. Haaf P, Garg P, Messroghli DR, Broadbent DA, Greenwood JP, Plein S. Cardiac T1 Mapping and Extracellular Volume (ECV) in clinical practice: a comprehensive review. J Cardiovasc Magn Reson. 2017 Jan;18(1):89.

10. Taylor AJ, Salerno M, Dharmakumar R, Jerosch-Herold M. T1 Mapping: Basic Techniques and Clinical Applications. JACC Cardiovasc Imaging. 2016 Jan;9(1):67–81.

11. Kim P, Hong Y, Im D, Suh YJ, Park C, Kim J, Chang S, Lee HJ, Hur J, Kim Y, Choi BW. Myocardial T1 and T2 Mapping: Techniques and Clinical Applications. Korean Journal of Radiology. 2017 Jan 1;18:113.

12. Menghoum N, Vos JL, Pouleur AC, Nijveldt R, Gerber BL. How to evaluate cardiomyopathies by cardiovascular magnetic resonance parametric mapping and late gadolinium enhancement. European Heart Journal - Cardiovascular Imaging. 2022 Apr 18;23(5):587–9.

13. Kuruvilla S, Adenaw N, Katwal AB, Lipinski MJ, Kramer CM, Salerno M. Late Gadolinium Enhancement on Cardiac Magnetic Resonance Predicts Adverse Cardiovascular Outcomes in Nonischemic Cardiomyopathy: A Systematic Review and Meta-Analysis. Circ: Cardiovascular Imaging. 2014 Mar;7(2):250–8.

14. Dyverfeldt P, Bissell M, Barker AJ, Bolger AF, Carlhäll CJ, Ebbers T, Francios CJ, Frydrychowicz A, Geiger J, Giese D, Hope MD, Kilner PJ, Kozerke S, Myerson S, Neubauer S, Wieben O, Markl M. 4D flow cardiovascular magnetic resonance consensus statement. J Cardiovasc Magn Reson. 2015 Dec;17(1):72.

15. Oscanoa JA, Middione MJ, Alkan C, Yurt M, Loecher M, Vasanawala SS, Ennis DB. Deep Learning-Based Reconstruction for Cardiac MRI: A Review. Bioengineering. 2023 Mar 6;10(3):334.

16. Ferreira PF, Gatehouse PD, Mohiaddin RH, Firmin DN. Cardiovascular magnetic resonance artefacts. Journal of Cardiovascular Magnetic Resonance. 2013 May 22;15(1):41.

17. Leiner T, Rueckert D, Suinesiaputra A, Baeßler B, Nezafat R, Išgum I, Young AA. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. J Cardiovasc Magn Reson. 2019 Oct 7;21(1):61.

18. Nacif MS, Zavodni A, Kawel N, Choi EY, Lima JAC, Bluemke DA. Cardiac magnetic resonance imaging and its electrocardiographs (ECG): tips and tricks. Int J Cardiovasc Imaging. 2012 Aug;28(6):1465–75.

19. Bhuva AN, Bai W, Lau C, Davies RH, Ye Y, Bulluck H, McAlindon E, Culotta V, Swoboda PP, Captur G, Treibel TA, Augusto JB, Knott KD, Seraphim A, Cole GD,

Petersen SE, Edwards NC, Greenwood JP, Bucciarelli-Ducci C, Hughes AD, Rueckert D, Moon JC, Manisty CH. A Multicenter, Scan-Rescan, Human and Machine Learning CMR Study to Test Generalizability and Precision in Imaging Biomarker Analysis. Circ: Cardiovascular Imaging. 2019 Oct;12(10):e009214.

20. Demir A, Wiesemann S, Erley J, Schmitter S, Trauzeddel RF, Pieske B, Hansmann J, Kelle S, Schulz-Menger J. Traveling Volunteers: A Multi-Vendor, Multi-Center Study on Reproducibility and Comparability of 4D Flow Derived Aortic Hemodynamics in Cardiovascular Magnetic Resonance. Magnetic Resonance Imaging. 2022 Jan;55(1):211–22.

21. Kranzusch R, aus dem Siepen F, Wiesemann S, Zange L, Jeuthe S, Ferreira da Silva T, Kuehne T, Pieske B, Tillmanns C, Friedrich MG, Schulz-Menger J, Messroghli DR. Z-score mapping for standardized analysis and reporting of cardiovascular magnetic resonance modified Look-Locker inversion recovery (MOLLI) T1 data: Normal behavior and validation in patients with amyloidosis. Journal of Cardiovascular Magnetic Resonance. 2020 Jan 20;22(1):6.

22. 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 Dec;22(1):17.

23. Bustin A, Fuin N, Botnar RM, Prieto C. From Compressed-Sensing to Artificial Intelligence-Based Cardiac MRI Reconstruction. Front Cardiovasc Med. 2020 Feb 25;7:17.

24. Kellman P, Hansen MS. T1-mapping in the heart: accuracy and precision. Journal of Cardiovascular Magnetic Resonance. 2014 Jan 4;16(1):2.

25. Hansen MS, Sørensen TS. Gadgetron: An open source framework for medical image reconstruction: Gadgetron. Magn Reson Med. 2013 Jun;69(6):1768–76.

26. Xue H, Inati S, Sørensen TS, Kellman P, Hansen MS. Distributed MRI reconstruction using gadgetron-based cloud computing: Gadgetron C-Bud Computing. Magn Reson Med. 2015 Mar;73(3):1015–25.

27. Marchesseau S, Ho JXM, Totman JJ. Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: A reproducibility study. European Journal of Radiology Open. 2016 Jan 1;3:60–6.

28. Raisi-Estabragh Z, Jaggi A, Gkontra P, McCracken C, Aung N, Munroe PB, Neubauer S, Harvey NC, Lekadir K, Petersen SE. Cardiac Magnetic Resonance Radiomics Reveal Differential Impact of Sex, Age, and Vascular Risk Factors on Cardiac Structure and Myocardial Tissue. Front Cardiovasc Med. 2021 Dec 22;8:763361.

29. Jha AK, Mithun S, Jaiswar V, Sherkhane UB, Purandare NC, Prabhash K, Rangarajan V, Dekker A, Wee L, Traverso A. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. Sci Rep. 2021 Jan 21;11(1):2055.

30. Avard E, Shiri I, Hajianfar G, Abdollahi H, Kalantari KR, Houshmand G, Kasani K, Bitarafan-rajabi A, Deevband MR, Oveisi M, Zaidi H. Non-contrast Cine Cardiac

Magnetic Resonance image radiomics features and machine learning algorithms for myocardial infarction detection. Computers in Biology and Medicine. 2022 Feb;141:105145.

31. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. Radiology. 2020 May;295(2):328–38.

32. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, Cetin I, Lekadir K, Camara O, Gonzalez Ballester MA, Sanroma G, Napel S, Petersen S, Tziritas G, Grinias E, Khened M, Kollerathu VA, Krishnamurthi G, Rohe MM, Pennec X, Sermesant M, Isensee F, Jager P, Maier-Hein KH, Full PM, Wolf I, Engelhardt S, Baumgartner CF, Koch LM, Wolterink JM, Isgum I, Jang Y, Hong Y, Patravali J, Jain S, Humbert O, Jodoin PM. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Trans Med Imaging. 2018 Nov;37(11):2514–25.

33. Campello VM, Gkontra P, Izquierdo C, Martín-Isla C, Sojoudi A, Full PM, Maier-Hein K, Zhang Y, He Z, Ma J, Parreño M, Albiol A, Kong F, Shadden SC, Acero JC, Sundaresan V, Saber M, Elattar M, Li H, Menze B, Khader F, Haarburger C, Scannell CM, Veta M, Carscadden A, Punithakumar K, Liu X, Tsaftaris SA, Huang X, Yang X, Li L, Zhuang X, Viladés D, Descalzo ML, Guala A, Mura LL, Friedrich MG, Garg R, Lebel J, Henriques F, Karakas M, Çavuş E, Petersen SE, Escalera S, Seguí S, Rodríguez-Palomares JF, Lekadir K. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. IEEE Transactions on Medical Imaging. 2021 Dec;40(12):3543–54.

34. Lalande A, Chen Z, Decourselle T, Qayyum A, Pommier T, Lorgis L, de la Rosa E, Cochet A, Cottin Y, Ginhac D, Salomon M, Couturier R, Meriaudeau F. Emidec: A Database Usable for the Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI. Data. 2020 Sep 24;5(4):89.

35. Mikami Y, Kolman L, Joncas SX, Stirrat J, Scholl D, Rajchl M, Lydell CP, Weeks SG, Howarth AG, White JA. Accuracy and reproducibility of semi-automated late gadolinium enhancement quantification techniques in patients with hypertrophic cardiomyopathy. J Cardiovasc Magn Reson. 2014 Dec;16(1):85.

36. Mariscal-Harana J, Kifle N, Razavi R, King AP, Ruijsink B, Puyol-Antón E. Improved AI-Based Segmentation of Apical and Basal Slices from Clinical Cine CMR. In: Puyol Antón E, Pop M, Martín-Isla C, Sermesant M, Suinesiaputra A, Camara O, Lekadir K, Young A, editors. Statistical Atlases and Computational Models of the Heart Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge. Cham: Springer International Publishing; 2022. p. 84–92.

37. Mullally J, Goyal P, Simprini LA, Afroz A, Kochav JD, Codella N, Devereux RB, Weinsaft JW. Marked variability in published CMR criteria for left ventricular basal slice selection - impact of methodological discrepancies on LV mass quantification. Journal of Cardiovascular Magnetic Resonance. 2013 Jan 30;15(1):P101.

38. Suinesiaputra A, Bluemke DA, Cowan BR, Friedrich MG, Kramer CM, Kwong R, Plein S, Schulz-Menger J, Westenberg JJM, Young AA, Nagel E. Quantification of

LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. J Cardiovasc Magn Reson. 2015 Jul 28;17:63.

39. Böttcher B, Lorbeer R, Stöcklein S, Beller E, Lang CI, Weber MA, Meinel FG. Global and Regional Test–Retest Reproducibility of Native T1 and T2 Mapping in Cardiac Magnetic Resonance Imaging. Journal of Magnetic Resonance Imaging. 2021;54(6):1763–72.

40. Gröschel J, Trauzeddel RF, Müller M, Von Knobelsdorff-Brenkenhoff F, Viezzer D, Hadler T, Blaszczyk E, Daud E, Schulz-Menger J. Multi-site comparison of parametric T1 and T2 mapping: healthy travelling volunteers in the Berlin research network for cardiovascular magnetic resonance (BER-CMR). J Cardiovasc Magn Reson. 2023 Aug 14;25(1):47.

41. Kim RJ, Simonetti OP, Westwood M, Kramer CM, Narang A, Friedrich MG, Powell AJ, Carr JC, Schulz-Menger J, Nagel E, Chan WS, Bremerich J, Ordovas KG, Rollings RC, Patel AR, Ferrari VA. Guidelines for training in cardiovascular magnetic resonance (CMR). Journal of Cardiovascular Magnetic Resonance. 2018 Aug 16;20(1):57.

42. Karamitsos TD, Hudsmith LE, Selvanayagam JB, Neubauer S, Francis JM. Operator induced variability in left ventricular measurements with cardiovascular magnetic resonance is improved after training. J Cardiovasc Magn Reson. 2007;9(5):777–83.

43. Raman SV, Markl M, Patel AR, Bryant J, Allen BD, Plein S, Seiberlich N. 30-minute CMR for common clinical indications: a Society for Cardiovascular Magnetic Resonance white paper. J Cardiovasc Magn Reson. 2022 Mar 1;24(1):13.

44. Ibrahim ESH, Frank L, Baruah D, Arpinar VE, Nencka AS, Koch KM, Muftuler LT, Unal O, Stojanovska J, Rubenstein JC, Brown SA, Charlson J, Gore EM, Bergom C. Value CMR: Towards a Comprehensive, Rapid, Cost-Effective Cardiovascular Magnetic Resonance Imaging. Bayford RH, editor. International Journal of Biomedical Imaging. 2021 May 15;2021:1–12.

45. Ibtehaz N, Rahman MS. MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. Neural Networks. 2020 Jan;121:74–87.

46. Dice LR, T. Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab; 1948.

47. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015 Dec;15(1):29.

48. Ammann C, Hadler T, Gröschel J, Kolbitsch C, Schulz-Menger J. Multilevel comparison of deep learning models for function quantification in cardiovascular magnetic resonance: On the redundancy of architectural variations. Frontiers in Cardiovascular Medicine [Internet]. 2023 [cited 2023 Jun 5];10. Available from: https://www.frontiersin.org/articles/10.3389/fcvm.2023.1118499

49. Sander J, de Vos BD, Wolterink JM, Išgum I. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. Medical Imaging 2019: Image Processing. 2019 Mar 15;44.

50. Sander J, de Vos BD, Išgum I. Automatic segmentation with detection of local segmentation failures in cardiac MRI. Sci Rep. 2020 Dec;10(1):21769.

51. Moon JC, Messroghli DR, Kellman P, Piechnik SK, Robson MD, Ugander M, Gatehouse PD, Arai AE, Friedrich MG, Neubauer S, Schulz-Menger J, Schelbert EB. Myocardial T1 mapping and extracellular volume quantification: a Society for Cardiovascular Magnetic Resonance (SCMR) and CMR Working Group of the European Society of Cardiology consensus statement. J Cardiovasc Magn Reson. 2013 Dec;15(1):92.

52. Hadler T, Wetzl J, Lange S, Geppert C, Fenski M, Abazi E, Gröschel J, Ammann C, Wenson F, Töpper A, Däuber S, Schulz-Menger J. Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging. Sci Rep. 2022 Dec;12(1):6629.

53. Hadler T, Ammann C, Wetzl J, Viezzer D, Gröschel J, Fenski M, Abazi E, Lange S, Hennemuth A, Schulz-Menger J. Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging. Computer Methods and Programs in Biomedicine. 2023 Aug;238:107615.

54. Mustra M, Delac K, Grgic M. Overview of the DICOM standard. In: 2008 50th International Symposium ELMAR. 2008. p. 39–44.

55. DICOM [Internet]. DICOM. [cited 2022 Mar 15]. Available from: https://www.dicomstandard.org

56. Mason D. SU-E-T-33: Pydicom: An Open Source DICOM Library. Medical Physics. 2011;38(6Part10):3493–3493.

57. Gillies S, others. Shapely: manipulation and analysis of geometric objects [Internet]. toblerity.org; 2007 [cited 2021 Dec 11]. Available from: https://github.com/Toblerity/Shapely

58. McKinney W. Data Structures for Statistical Computing in Python. In Austin, Texas; 2010 [cited 2023 Nov 14]. p. 56–61. Available from: https://conference.scipy.org/proceedings/scipy2010/mckinney.html

59. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering. 2007 May;9(3):90–5.

60. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021 Apr 6;6(60):3021.

61. Qt 5.15 [Internet]. [cited 2021 Dec 11]. Available from: https://doc.qt.io/qt-5/

62. PyQt - QTab Widget [Internet]. [cited 2022 Mar 15]. Available from: https://www.tutorialspoint.com/pyqt/pyqt_qtabwidget.htm

63. matplotlib.figure.Figure — Matplotlib 3.3.4 documentation [Internet]. [cited 2022 Mar 15]. Available from: https://matplotlib.org/3.3.4/api/_as_gen/matplotlib.figure.Figure.html

64. pandas.DataFrame — pandas 1.4.1 documentation [Internet]. [cited 2022 Mar 15]. Available from: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

65. Cerqueira MD, Weissman NJ, Dilsizian V, Jacobs AK, Verani MS. Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart. Journal of the American Heart Association. :4.

66. Gillies S, others. Rasterio: geospatial raster I/O for Python programmers [Internet]. Mapbox; 2013 [cited 2021 Dec 11]. Available from: https://github.com/mapbox/rasterio

67. Fenski M, Grandy TH, Viezzer D, Kertusha S, Schmidt M, Forman C, Schulz-Menger J. Isotropic 3D compressed sensing (CS) based sequence is comparable to 2D-LGE in left ventricular scar quantification in different disease entities. Int J Cardiovasc Imaging. 2022 Aug 1;38(8):1837–50.

68. Gröschel J, Ammann C, Zange L, Viezzer D, Forman C, Schmidt M, Blaszczyk E, Schulz-Menger J. Fast acquisition of left and right ventricular function parameters applying cardiovascular magnetic resonance in clinical routine – validation of a 2-shot compressed sensing cine sequence. Scandinavian Cardiovascular Journal. 2022 Dec 31;56(1):266–75.

69. 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 Sep;7(9):882–92.

70. Kawel-Boehm N, Hetzel SJ, Ambale-Venkatesh B, Captur G, Francois CJ, Jerosch-Herold M, Salerno M, Teague SD, Valsangiacomo-Buechel E, van der Geest RJ, Bluemke DA. Reference ranges (“normal values”) for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. Journal of Cardiovascular Magnetic Resonance. 2020 Dec 14;22(1):87.

71. Luu JM, Gebhard C, Ramasundarahettige C, Desai D, Schulze K, Marcotte F, Awadalla P, Broet P, Dummer T, Hicks J, Larose E, Moody A, Smith EE, Tardif JC, Teixeira T, Teo KK, Vena J, Lee DS, Anand SS, Friedrich MG, the CAHHM Study Investigators. Normal sex and age-specific parameters in a multi-ethnic population: a cardiovascular magnetic resonance study of the Canadian Alliance for Healthy Hearts and Minds cohort. J Cardiovasc Magn Reson. 2022 Dec;24(1):2.

72. Full PM, Isensee F, Jäger PF, Maier-Hein K. Studying Robustness of Semantic Segmentation Under Domain Shift in Cardiac MRI. In: Puyol Anton E, Pop M, Sermesant M, Campello V, Lalande A, Lekadir K, Suinesiaputra A, Camara O, Young A, editors. Statistical Atlases and Computational Models of the Heart M&Ms and EMIDEC Challenges [Internet]. Cham: Springer International Publishing; 2021 [cited 2023 Jul 15].

p. 238–49. (Lecture Notes in Computer Science; vol. 12592). Available from: https://link.springer.com/10.1007/978-3-030-68107-4_24

73. Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI. The MIDAS Journal [Internet]. 2009 Jul 9 [cited 2023 Jul 7]; Available from: https://www.midasjournal.org/browse/publication/658

74. Petitjean C, Zuluaga MA, Bai W, Dacher JN, Grosgeorge D, Caudron J, Ruan S, Ayed IB, Cardoso MJ, Chen HC, Jimenez-Carretero D, Ledesma-Carbayo MJ, Davatzikos C, Doshi J, Erus G, Maier OMO, Nambakhsh CMS, Ou Y, Ourselin S, Peng CW, Peters NS, Peters TM, Rajchl M, Rueckert D, Santos A, Shi W, Wang CW, Wang H, Yuan J. Right ventricle segmentation from cardiac MRI: A collation study. Medical Image Analysis. 2015 Jan 1;19(1):187–202.

75. Food and Drug Administration (FDA), Health Canada, Medicines and Healthcare Products Regulatory Agency (MHRA)). Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021.

76. Riazy L, Daeuber S, Lange S, Viezzer D, Ott S, Wiesemann S, Blaszczyk E, Mühlberg F, Zange L, Schulz-Menger J. Translating principles of quality control to cardiovascular magnetic resonance: assessing quantitative parameters of the left ventricle in a large cohort. Scientific Reports. 2023 Feb 7;13.

77. Shwartzman O, Gazit H, Shelef I, Riklin-Raviv T. The Worrisome Impact of an Inter-rater Bias on Neural Network Training. arXiv:190611872 [cs, eess] [Internet]. 2020 May 31 [cited 2021 Dec 11]; Available from: http://arxiv.org/abs/1906.11872

78. Shivaprasad D. Bigdata: A Survey On RDBMS And Various NOSQL Databases On Storing Medical Images [Internet]. [cited 2023 Jun 3]. Available from: https://www.ijarnd.com/manuscripts/v2i5/V2I5-1156.pdf

79. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, Lyu J. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Military Medical Research. 2021 Aug 11;8(1):44.

80. Margeta J. Machine Learning for Simplifying the Use of Cardiac Image Databases. Signal and Image Processing. 2015;194.

81. Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. Food and Drug Administration, HHS; 2019.

82. Vergani V, Razavi R, Puyol-Antón E, Ruijsink B. Deep Learning for Classification and Selection of Cine CMR Images to Achieve Fully Automated Quality-Controlled CMR Analysis From Scanner to Report. Front Cardiovasc Med. 2021 Oct 14;8:742640.

83. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017 May 24;60(6):84–90.

84. Rajchl M, Lee MCH, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Damodaram M, Rutherford MA, Hajnal JV, Kainz B, Rueckert D. DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. IEEE Trans Med Imaging. 2017 Feb;36(2):674–83.

85. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017 Apr;39(4):640–51.

86. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41. (Lecture Notes in Computer Science).

87. 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 Feb;18(2):203–11.

88. Chen C, Biffi C, Tarroni G, Petersen S, Bai W, Rueckert D. Learning Shape Priors for Robust Cardiac MR Segmentation from Multi-view Images. In 2019 [cited 2022 Jul 21]. p. 523–31. Available from: http://arxiv.org/abs/1907.09983

89. Zotti C, Luo Z, Lalande A, Jodoin PM. Convolutional Neural Network With Shape Prior Applied to Cardiac MRI Segmentation. IEEE J Biomed Health Inform. 2019 May;23(3):1119–28.

90. Multi-sequence myocardium segmentation with cross-constrained shape and neural network-based initialization. Computerized Medical Imaging and Graphics. 2019 Jan 1;71:49–57.

91. Duan J, Bello G, Schlemper J, Bai W, Dawes TJW, Biffi C, de Marvao A, Doumoud G, O'Regan DP, Rueckert D. Automatic 3D Bi-Ventricular Segmentation of Cardiac Images by a Shape-Refined Multi- Task Deep Learning Approach. IEEE Trans Med Imaging. 2019 Sep;38(9):2151–64.

92. Painchaud N, Skandarani Y, Judge T, Bernard O, Lalande A, Jodoin PM. Cardiac MRI Segmentation with Strong Anatomical Guarantees. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap PT, Khan A, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 [Internet]. Cham: Springer International Publishing; 2019 [cited 2023 Jul 9]. p. 632–40. (Lecture Notes in Computer Science; vol. 11765). Available from: https://link.springer.com/10.1007/978-3-030-32245-8_70

93. Dechesne C, Lassalle P, Lefevre S. Bayesian Deep Learning with Monte Carlo Dropout for Qualification of Semantic Segmentation. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS [Internet]. Brussels, Belgium: IEEE; 2021 [cited 2022 Sep 16]. p. 2536–9. Available from: https://ieeexplore.ieee.org/document/9555043/

94. Valindria VV, Lavdas I, Bai W, Kamnitsas K, Aboagye EO, Rockall AG, Rueckert D, Glocker B. Reverse Classification Accuracy: Predicting Segmentation Performance

in the Absence of Ground Truth. IEEE Trans Med Imaging. 2017 Aug;36(8):1597–606.

95. Hann E, Popescu IA, Zhang Q, Gonzales RA, Barutçu A, Neubauer S, Ferreira VM, Piechnik SK. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Med Image Anal. 2021 Jul;71:102029.

96. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis. 2020 Feb;128(2):336–59.

97. Gröschel J, Kuhnt J, Viezzer D, Hadler T, Hormes S, Barckow P, Schulz-Menger J, Blaszczyk E. Comparison of manual and artificial intelligence based quantification of myocardial strain by feature tracking—a cardiovascular MR study in health and disease. Eur Radiol [Internet]. 2023 Aug 18 [cited 2023 Oct 8]; Available from: https://link.springer.com/10.1007/s00330-023-10127-y

98. Bello GA, Dawes TJW, Duan J, Biffi C, de Marvao A, Howard LSGE, Gibbs JSR, Wilkins MR, Cook SA, Rueckert D, O'Regan DP. Deep learning cardiac motion analysis for human survival prediction. Nat Mach Intell. 2019 Feb 11;1:95–104.

Statutory Declaration

“I, Thomas Hadler, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic:
„A Multilevel Reader Comparison Software Tool for Semi-Automated Quality Control in Cardiovascular Magnetic Resonance Imaging: Lazy Luna / Eine mehrstufige Auswerter-Vergleichssoftware für halbautomatisierte Qualitätskontrolle in der kardiovaskulären Magnetresonanztomographie: Lazy Luna”,
independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts, which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; http://www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

Date Signature

Declaration of your own Contribution to the Publications

Thomas Hadler contributed the following to the below listed publications:

Publication 1: **Hadler, T**., Wetzl, J., Lange, S., Geppert, C., Fenski, M., Abazi, E., Gröschel, J., Amman, C., Wenson, F., Töpper, A., Däuber, S., Schulz-Menger, J. "Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging", Scientific Reports, 2022
Journal Impact Factor 2021: 4,996
Contribution:

- Idea of software
- Literature research
- Conception of software (Design of data interfaces for DICOM images, exact annotation representations, sorting and accessibility of data, data storage system, backend and frontend availability for multiple users, programing language choice, and assurance of operating system independence)
- Implementation of software and data interface
- Compilation of data (excluding the generation of images and contours, which was performed by dedicated medical personal)
- Use of Lazy Luna (the implemented reader comparison software) to compare two expert readers on short-axis cine quantification (i.e. calculation of clinical parameters and their differences, segmentation metrics for contour types, and tracing reader biases to qualitative example images)
- Creation of manuscript, all figures and all tables
- Discussion with coauthors
- Revision of manuscript, figures and tables
- Presentation of results at congresses

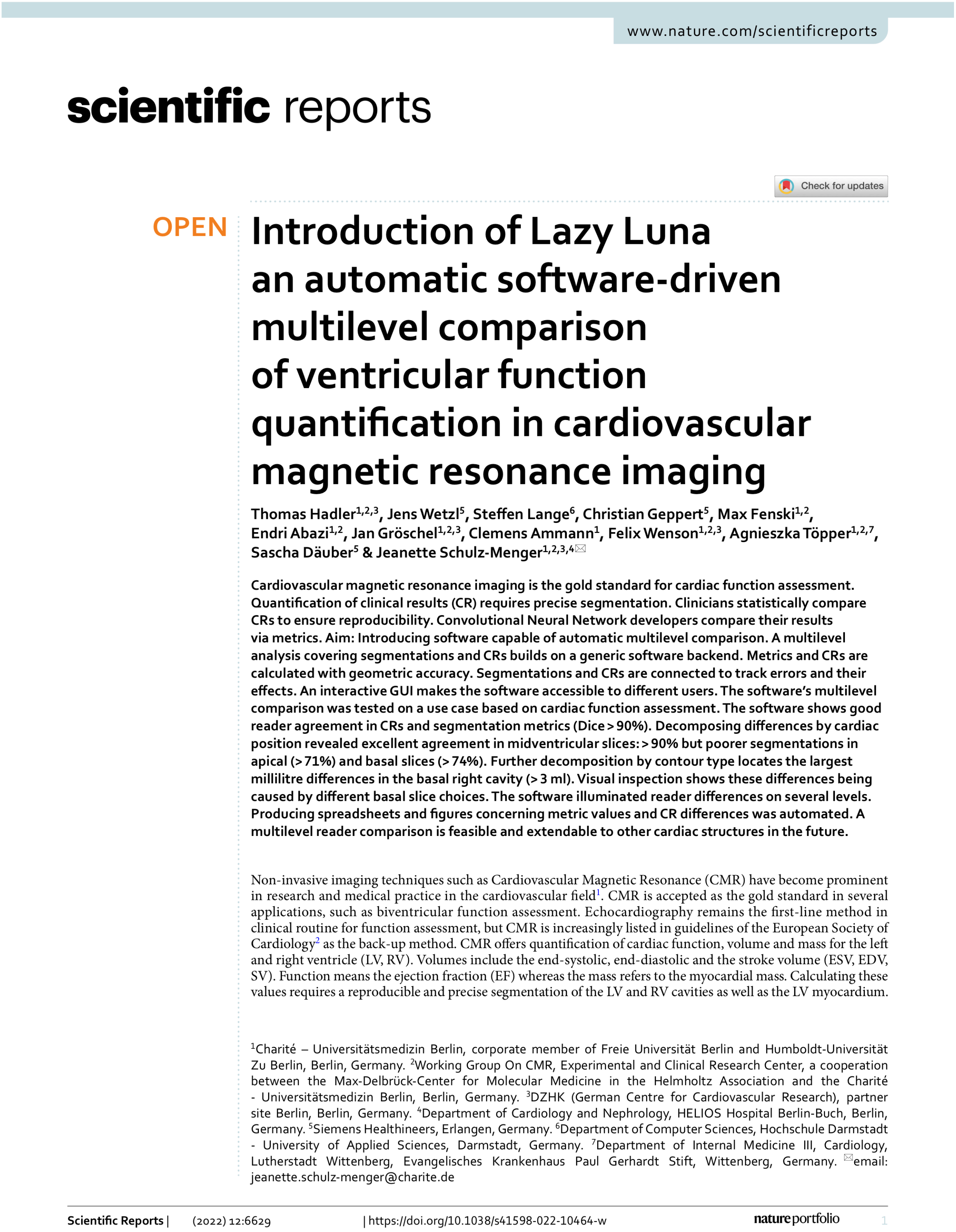
Publication 2: **Hadler, T**., Ammann, C., Wetzl, J., Viezzer, D., Gröschel, J., Fenski, M., Abazi, E., Lange, S., Hennemuth, A., Schulz-Menger, J. "Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging", Computer Methods and Programs in Biomedicine, 2023
Journal Impact Factor 2021: 7,027
Contribution:

- Idea of software
- Literature research
- Conception of software extendibility (Redesign of data interfaces for DICOM images, exact annotation representations in a generic fashion, extendibility of interface classes for figures, tables and other graphical user interface elements)
- Implementation of extendible software
- Compilation of data (excluding the generation of images and contours, which was performed by dedicated medical personal, or acquired from an openly available segmentation competition)

- Use of Lazy Luna to compare two experts on parametric T1 mapping data and an expert and an artificial intelligence on Late Gadolinium Enhancement data (i.e. calculation of clinical parameters and their differences, segmentation metrics for contour types, and tracing reader biases to qualitative example images)
- Creation of manuscript, all figures and all tables
- Discussion with coauthors
- Revision of manuscript, figures and tables

Signature, date and stamp of first supervising university professor / lecturer

Signature of doctoral candidate



www.nature.com/scientificreports

scientific reports

OPEN

Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging

Thomas Hadler[1,2,3], Jens Wetzl[5], Steffen Lange[6], Christian Geppert[5], Max Fenski[1,2], Endri Abazi[1,2], Jan Gröschel[1,2,3], Clemens Ammann[1], Felix Wenson[1,2,3], Agnieszka Töpper[1,2,7], Sascha Däuber[5] & Jeanette Schulz-Menger[1,2,3,4]✉

Cardiovascular magnetic resonance imaging is the gold standard for cardiac function assessment. Quantification of clinical results (CR) requires precise segmentation. Clinicians statistically compare CRs to ensure reproducibility. Convolutional Neural Network developers compare their results via metrics. Aim: Introducing software capable of automatic multilevel comparison. A multilevel analysis covering segmentations and CRs builds on a generic software backend. Metrics and CRs are calculated with geometric accuracy. Segmentations and CRs are connected to track errors and their effects. An interactive GUI makes the software accessible to different users. The software's multilevel comparison was tested on a use case based on cardiac function assessment. The software shows good reader agreement in CRs and segmentation metrics (Dice > 90%). Decomposing differences by cardiac position revealed excellent agreement in midventricular slices: > 90% but poorer segmentations in apical (> 71%) and basal slices (> 74%). Further decomposition by contour type locates the largest millilitre differences in the basal right cavity (> 3 ml). Visual inspection shows these differences being caused by different basal slice choices. The software illuminated reader differences on several levels. Producing spreadsheets and figures concerning metric values and CR differences was automated. A multilevel reader comparison is feasible and extendable to other cardiac structures in the future.

Non-invasive imaging techniques such as Cardiovascular Magnetic Resonance (CMR) have become prominent in research and medical practice in the cardiovascular field[1]. CMR is accepted as the gold standard in several applications, such as biventricular function assessment. Echocardiography remains the first-line method in clinical routine for function assessment, but CMR is increasingly listed in guidelines of the European Society of Cardiology[2] as the back-up method. CMR offers quantification of cardiac function, volume and mass for the left and right ventricle (LV, RV). Volumes include the end-systolic, end-diastolic and the stroke volume (ESV, EDV, SV). Function means the ejection fraction (EF) whereas the mass refers to the myocardial mass. Calculating these values requires a reproducible and precise segmentation of the LV and RV cavities as well as the LV myocardium.

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität Zu Berlin, Berlin, Germany. [2]Working Group On CMR, Experimental and Clinical Research Center, a cooperation between the Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association and the Charité - Universitätsmedizin Berlin, Berlin, Germany. [3]DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany. [4]Department of Cardiology and Nephrology, HELIOS Hospital Berlin-Buch, Berlin, Germany. [5]Siemens Healthineers, Erlangen, Germany. [6]Department of Computer Sciences, Hochschule Darmstadt - University of Applied Sciences, Darmstadt, Germany. [7]Department of Internal Medicine III, Cardiology, Lutherstadt Wittenberg, Evangelisches Krankenhaus Paul Gerhardt Stift, Wittenberg, Germany. ✉email: jeanette.schulz-menger@charite.de

In clinical practice as well as in research, readers annotate contours often in accordance with the SCMR guidelines[1]. However, manual segmentation is time-consuming and remains prone to inter- and intraobserver variability[3,4]. In order to characterize pathologies with diagnostic approaches, inter- and intraobserver analyses are performed in order to ensure the methods' statistical reproducibility and accuracy[3–6]. Segmentations are based on subpixel resolution producing contours as polygons[1]. An objective analysis of segmentation differences could be based on segmentation metrics such as the Dice Similarity Coefficient (Dice) or the Hausdorff Distance (HD) as typically used in computer vision challenges and tasks[7–9]. Metrics are typically not used to compare segmentation similarity in context to clinical relevance and decision-making.

In recent years several convolutional neural network (CNN) developers have trained CNNs to contour CMR-images similar to medical experts[9–14]. The annotations are generated in a fraction of the time a reader would require and are often performed on subpixel resolution as segmentation masks[9,13–15]. CNNs demonstrate promising clinical results within the variability of interobserver errors[16,17], while still making human atypical mistakes[18–20]. Segmentation metrics (such as the Dice and HD) are typically used to compare CNNs to medical readers on the level of individual segmentations[9,16,21]. The qualitative nature of the human atypical segmentation differences remains elusive[18,20].

The goal of this paper is to design software that is capable of an automatic multilevel reader comparison. Usability by CNN developers as backend software and by medical experts as a graphical user interface (GUI) should be given alike.

Methods

The software Lazy Luna was designed to offer a multilevel reader comparison that covers segmentations and CRs. Metrics and CRs are calculated accurately. Segmentations and CRs are connected to allow for error tracking. An interactive GUI makes the software accessible to clinical readers and CNN developers. Lazy Luna's functionality was demonstrated by performing a multilevel interobserver analysis.

Data.

The dataset encompasses short-axis balanced steady-state free precession (bSSFP) cine CMR images of 13 patients (39 ± 13 years, 7/6 male/female). They were produced on a 1.5 T Avanto fit, Siemens Healthineers. The cases were selected randomly from an on-going trial. The central criterion was the performance of an interobserver analysis of the right and left ventricle. A short image stack consists of 16–18 slices and 30 phases. Two expert readers segmented the images using Circle Cardiovascular Imaging: cvi42 version 5.12.1.[22]. They segmented the LV and RV cavity and contoured the LV myocardium and papillary muscles.

The local ethics committee of Charité Medical University Berlin gave ethics approval for the original study (approval number EA1/198/20). All patients gave their written informed consent before participating in the study. All methods were carried out in accordance with relevant guidelines and regulations.

Cases.

Cases contain images, annotations (i.e. segmentations, points, etc.) of these images and clinical values that were calculated on the basis of these images and their annotations (Fig. 1a). The images were sorted into phases and slices. Two cases can be compared to each other when they reference the same original images. When many comparable cases were segmented by two readers statistics can be performed on the metric values and CRs (Fig. 1).

The images, segmentations and CRs refer to the same case allowing for tracking the effect of failed segmentations on differences in assessed CRs. For many comparable cases the outliers of CRs can be identified and the causes for their particularity backtracked to their origin in specific contours and their cardiac position (i.e. basal, midventricular, apical).

The images are stored in the Dicom[23] (Digital Imaging and Communications in Medicine) format. Dicom images are used to store images as well as information pertaining to those images. The images are loaded using the Python package Pydicom[24]. Annotations are stored in a custom Lazy Luna format, as pickle files containing a Python dictionary that maps contour names to Shapely[25] objects. Shapely is described in "Geometrical representation and metrics".

Data pre-processing.

Lazy Luna was designed to emphasize precision. The analysis tool can only be applied if the user transforms images and annotations to fit Lazy Luna's interface. Lazy Luna requires images in Dicom format and annotations as pickle-files containing Shapely objects. Thus, pre-processing the data is a requirement for using the tool. An easy to use Data Pre-processing GUI for labelling Dicom images as well as linking the images to segmentations was used.

Finding the short-axis cine Dicom images in a set of several thousand images is an error-prone task and user intervention is essential. Images are manually identified as short-axis cine images by adding a Lazy Luna Dicom tag. The clinicians contoured the relevant images and stored the contours as workspaces. These workspaces were converted into the custom Lazy Luna annotation format.

Geometrical representation and metrics.

Lazy Luna uses Shapely to process annotations. Shapely is a Python package for manipulating and analysing geometric objects (i.e. polygons, lines, points)[25]. Segmentations are modelled as polygons (LV, RV endocardial contour and LV myocardium) or MultiPolygons (papillary muscles). Shapely is capable of performing a wide array of precise geometrical operations, such as area calculation, intersection, union and calculating the Hausdorff distance (HD)[26]. The Dice metric is calculated using intersection and union operations on two Shapely objects (Fig. 1b). The millilitres and their differences (ml Diff) are calculated using Dicom tag information on pixel height, width and slice thickness in mm:

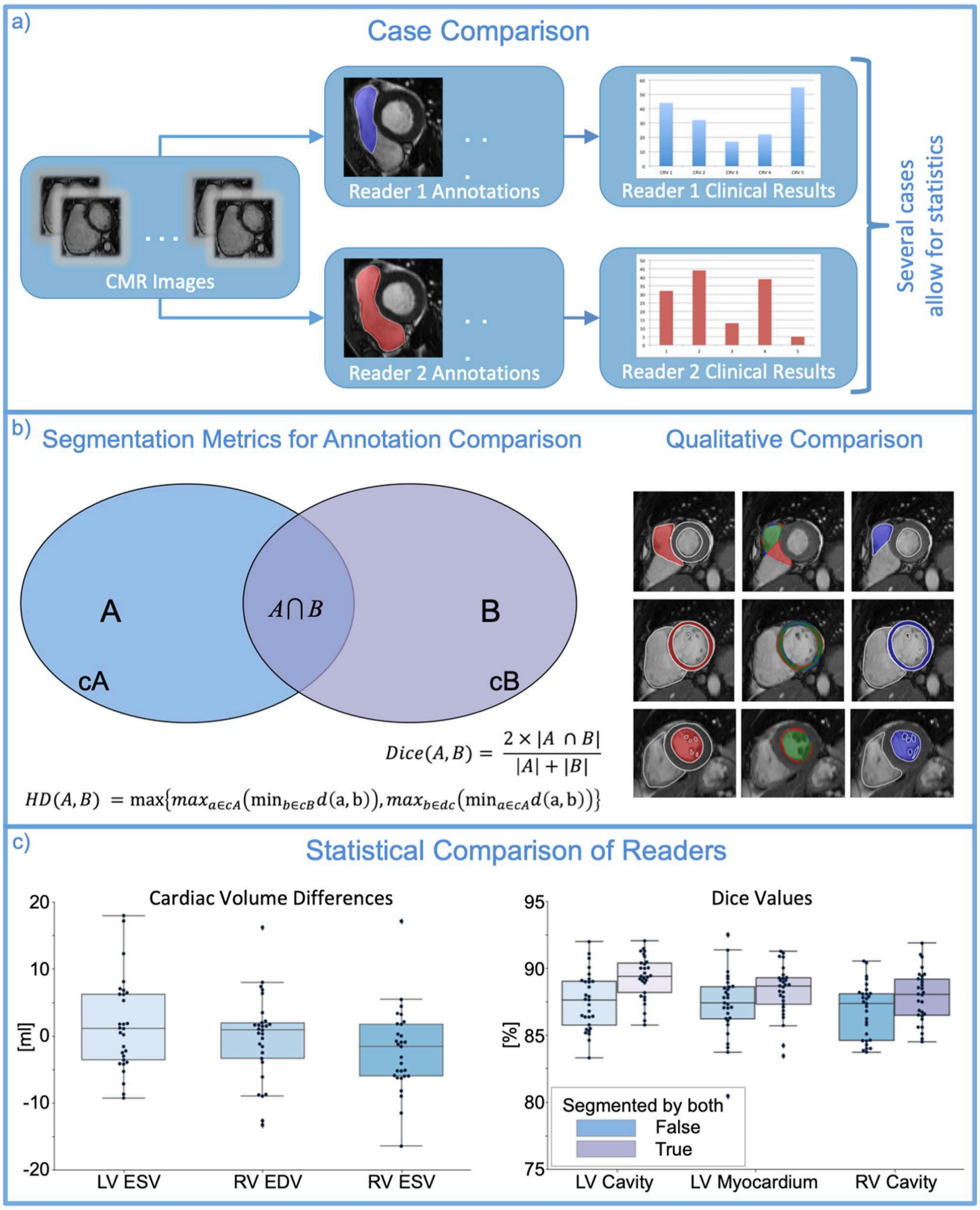

a)
Case Comparison
CMR Images
Reader 1 Annotations
Reader 2 Annotations
Reader 1 Clinical Results
Reader 2 Clinical Results
Several cases allow for statistics
b)
Segmentation Metrics for Annotation Comparison
Qualitative Comparison
A
B
$A \cap B$
cA
cB
$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$
$HD(A,B) = \max\{max_{a\in cA}(\min_{b\in cB} d(a,b)), max_{b\in dc}(\min_{a\in cA} d(a,b))\}$
c)
Statistical Comparison of Readers
Cardiac Volume Differences
[ml]
20
10
0
-10
-20
LV ESV
RV EDV
RV ESV
Dice Values
[%]
95
90
85
80
75
LV Cavity
LV Myocardium
RV Cavity
Segmented by both
False
True

Figure 1. Multilevel Reader Comparison. Caption: At the top (**a**) a case comparison is presented. Comparable cases concern CMR images that were segmented by two different readers. Clinical results are generated from images and segmentations. Segmentation metrics (**b**), such as Dice and Hausdorff metric, provide quantitative comparisons of segmentations. Next to the metrics, qualitative visualizations of segmentation differences are presented. The first reader's segmentation is coloured blue, the second red and their agreement in green. Reader comparisons are modelled as the distributions of clinical result differences and metric value distributions. LV: Left ventricle, RV: Right ventricle, ESV: End-systolic volume, EDV: End-diastolic volume, cA (cB): Contour of Area A, ml: Millilitre, HD: Hausdorff metric.

$$ml\,Diff\,(A,B) = (|A| - |B|) \times \; area\,per\,pixel \; \times \; slice\,thickness$$

$$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

$$HD(A,B) = \max\{max_{a \in cA}(\min_{b \in cB} d(\mathrm{a},\mathrm{b})), max_{b \in dc}(\min_{a \in cA} d(\mathrm{a},\mathrm{b}))\}$$

We offer two different averages for the Dice metric. The first one averages over all images, the second only over images segmented by both readers. The first rewards correct segmentation decisions, e.g. if the CNN should not and does not segment an image it considers this as an example of 100% Dice. If it makes an incorrect segmentation decision then it considers this mistake as 0% Dice. The second Dice average only considers the segmentation similarity for segmented images and discounts the relevance of the segmentation decision. It exclusively reflects the similarity of segmentation areas.

In order to calculate precise values for segmentation masks (typical outputs of CNNs) these must also be converted to Shapely objects. The transformation method should outline the pixelated segmentation mask precisely. For example, Rasterio's rasterize function can be used to produce outlines of segmentation masks in Shapely format[27].

Software conception. The software Lazy Luna builds on several implemented classes following the object oriented programming paradigm. Classes are indicated with a capital letter. The Cases described above are a container class for images and annotations. An Annotation Type (i.e. segmentations of short-axis cine images) can be attached to a case and offers several visualization functions as well as geometric operations. Categories can be attached to a case in order to structure the case's images into slices and phases by using Dicom image information. Categories identify relevant phases for Clinical Results. Clinical Result classes can be attached to a Case in order to calculate CRs based on the images, annotations and categories. Case Comparisons contain two cases that reference the same images. Metrics can be attached to a Case Comparison to calculate metric values.

Figures are classes that inherit their behaviour from the Python package Matplotlib[28]. Matplotlib figures allow for creating professional static and interactive visualizations. Seaborn[29] (a wrapper Python package around Matplotlib) is used for statistical visualizations (Fig. 2). Tables are classes that extend Pandas DataFrame objects. Pandas[30] allows for extensive data analysis and easy storing of spreadsheets, extensive tabular information transformation and data manipulation.

The graphical user interface (GUI) builds on PyQt5, which has Python bindings to Qt version 5[31]. Matplotlib figures and DataFrames are easy to integrate into PyQt5 GUIs. Interactive Matplotlib figures (Figs. 3, 4) can also be integrated, allowing for tracking function by linking different figures to each other that offer insights on several levels of analysis (such as CRs and metric values, or metric values and qualitative visualizations).

Lazy Luna offers several automated outputs. These include the calculation of tables of metric values (for all phases and slices) for all cases and the calculation of tables of CRs and their differences for all cases (supplementary information). It also produces summary tables for clinical value differences and a metric evaluation of the contours they are based on (Table 1) and for the metric values decomposed by contour type and cardiac position (Table 2). Lazy Luna offers the automatic generation of figures, such as Bland–Altman plots for clinical value distributions and Dice values as boxplots (Fig. 2).

Ethical approval. The local ethics committee of Charité Medical University Berlin gave ethics approval for the original study (approval number EA1/323/15). All patients gave their written informed consent before participating in the study. All methods were carried out in accordance with relevant guidelines and regulations.

Results

It was possible and feasible to merge the evaluation methods of medical experts and CNN developers. The software automatically structures Dicom images and annotations allowing for comparisons between readers. The cases are compared via their segmentations and CR simultaneously while tracking errors. Calculating all metrics and CRs on the contour level provides sub-pixel accuracy. Lazy Luna can be used to perform inter- and intraobserver analyses. As the software package is described in "Methods" the results section presents Lazy Luna's GUI and its generated outputs to illustrate a reader comparison performed with Lazy Luna.

Quantitative results for the use-case. The comparison of the readers' cardiac function assessments produced the following analysis. The readers show good general agreement on quantitative CRs and segmentation metric values (Table 1). Lazy Luna calculated a CRs spreadsheet (supplementary information), which was used to calculate Pearson's correlation coefficients for the CRs assessed by both readers. These are LVESV: 91%, LVEDV: 99%, RVESV: 96%, RVEDV: 95%, LVSV: 95%, LVEF: 74%, RVSV 87%, RVEF: 78%, LVM: 97%. Average Dice values are 91.9% for all images and 92.2% for images segmented by both readers. Details are in Table 1. Furthermore, these results can be displayed as single plots to illustrate the result similarities and differences. This is given in Fig. 2, which shows an automatically produced overview of CRs.

Qualitative results for the use-case. Furthermore, the use-case was also evaluated qualitatively with a visualization of segmentation differences, which was implemented for the GUI. That allows an identification of

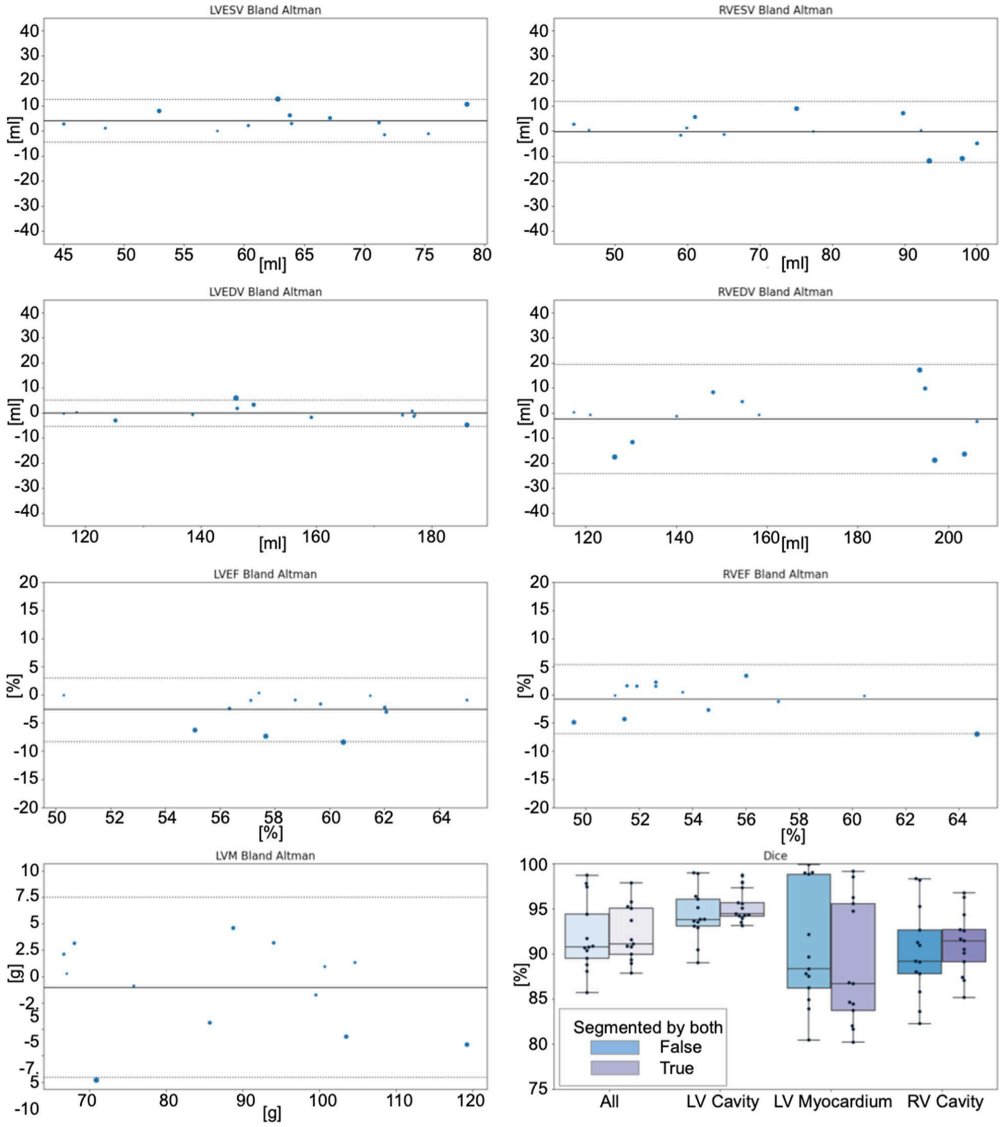


Figure 2. Automatic Generation of Clinical Results Overview. Caption: This plot is automatically generating after loading the cases into Lazy Luna's GUI. Bland-Altman plots show clinical result averages and differences as points for all cases. Point size represents difference size. The solid line marks the mean difference between readers; the dashed lines mark the mean differences ±1.96 standard deviations. The last plot offers two Dice boxplots per contour type, one for all images, another restricted to images segmented by both readers. The clinical result differences hover around zero for the LV and the RV. The variance is larger for results concerning the RV. Dice values are higher for the LV cavity than for LV myocardium or RV cavity. GUI: Graphical user interface, RV: Right ventricle, LV: Left ventricle, ESV: End-systolic volume, EDV: End-diastolic volume, EF: Ejection fraction, LVM: Left ventricular mass, Dice: Dice similarity coefficient.

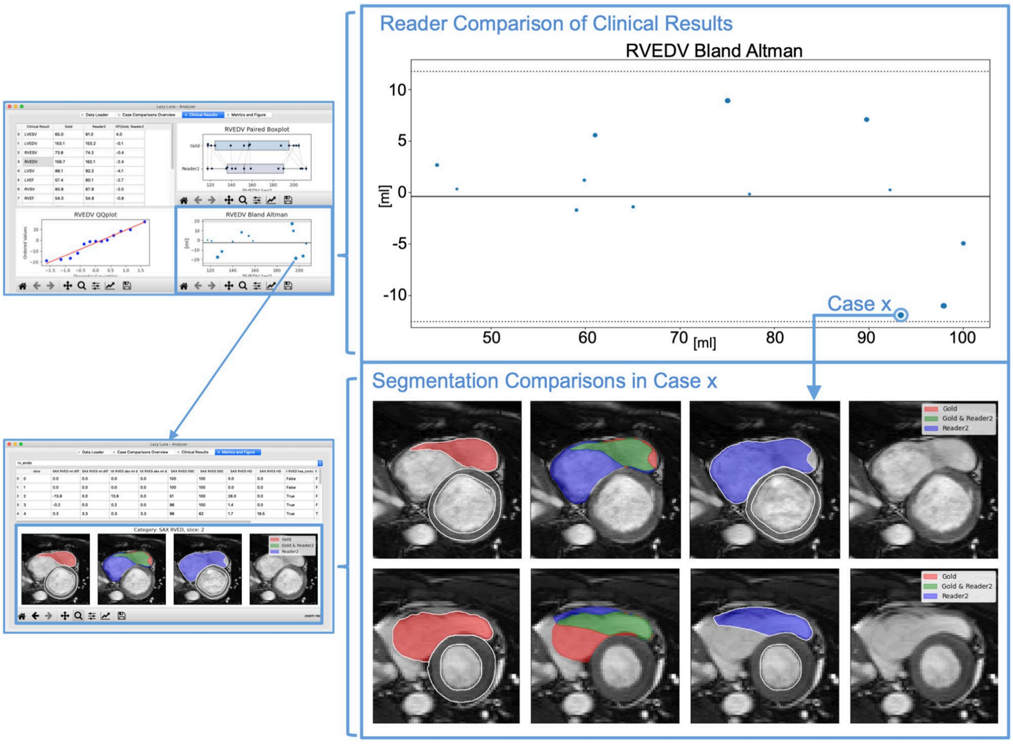


Figure 3. Tracking and Visualizing Reader Differences with Lazy Luna's GUI. Caption: On the left, two tabs of Lazy Luna's GUI are presented. On the right, parts of these tabs are magnified. For the top tab the RVESV Bland-Altman plot (outlined in blue) is magnified. For the bottom tab the visualization of segmentation differences is magnified. The first reader's segmentations (subplot 1) are in red, the second reader's are in blue (subplot 3). In the second subplot agreement is in green and areas exclusive to one reader are in that reader's respective colour. The top tab includes a table of clinical result averages per reader next to their average differences (top left), a QQ-plot (bottom left) and paired boxplots (top right). Clicking a point in the Bland-Altman plot opened the lower tab. This tab's table presents all metric values concerning the case's segmentations. RV: Right ventricle, ESV: End-systolic volume, GUI: Graphical user interface, QQ-plot: Quantile-quantile plot.

different slice selection or interpretation, which may lead to large volume differences. An example of a disagreement is given in Fig. 3.

Tracking differences in the use-case. CR differences can be caused in different cardiac positions and structures. Lazy Luna can track segmentation differences and their impacts on CRs. For this use-case investigating the cardiac position of segmentation difficulties reveals that the midventricular slices have higher Dice values for all contour types (LV cavity: 97%, LV myocardium: 91%, RV cavity: 94%). As a result millilitre differences remain small in these slices (< 1 ml). Segmentation difficulties are larger in basal and apical slices (Table 2). The Dice metric is poorest for the LV myocardium in the apical slices (74%). However, the impact in clinical values is smaller because the millilitre differences remain small (< 0.5 ml). The Dice metric values are also lower in the basal slices (LV cavity: 88%, LV myocardium: 87%, RV cavity: 72%) (Table 2, Fig. 3). However, the millilitre differences are larger in the basal slices, especially those concerning the RV (> 3 ml, Table 2), which causes larger millilitre differences in the CRs. One of Lazy Luna's interactive GUI tabs allows for exploring this phenomenon (Fig. 4). An interactive metrics correlation plot shows that RV endocardial segmentation disagreements produce the largest RV millilitre differences and provides visualizations of selected differences.

Discussion

Our main achievement is the implementation of the investigative software Lazy Luna, which is capable of performing a multilevel analysis on reader differences with a graphical user interface. The functionality of Lazy Luna was illustrated by carrying out an interobserver analysis between two experienced readers. This analysis

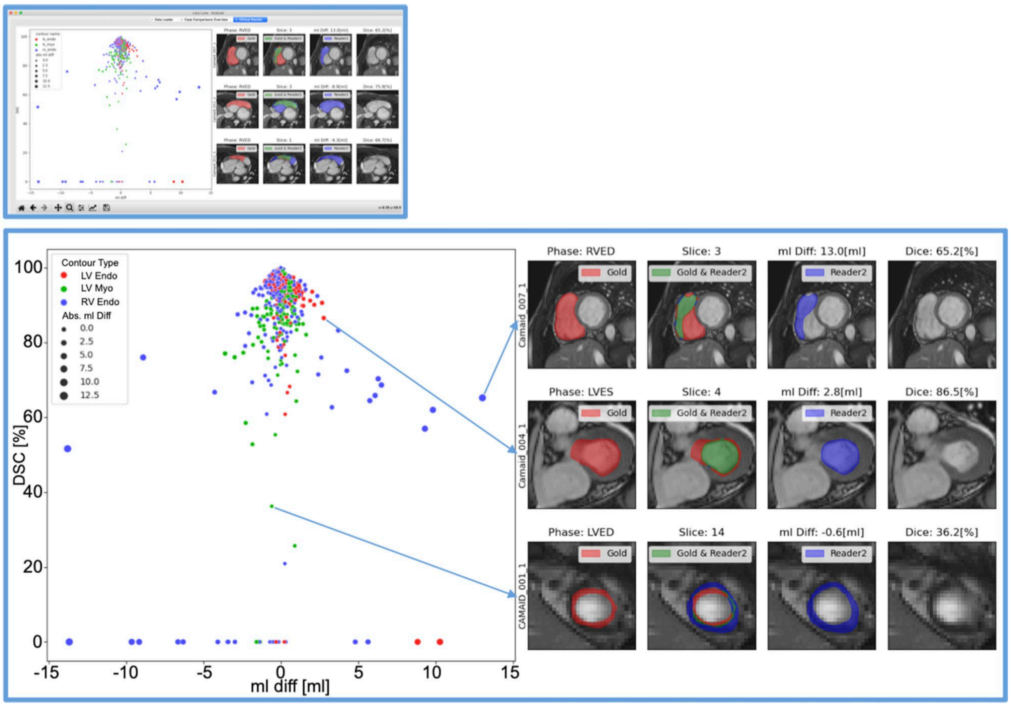


Figure 4. Interactive Correlation Plots of Segmentation Comparisons. Caption: The above window shows the interactive plot in the GUI. Below the plot is enlarged. Every point represents a contour comparison as millilitre difference and Dice value. Its colour distinguishes LV endocardial contour (red), LV myocardium (green) and RV endocardial contour (blue) contours. The point size represents the absolute millilitre difference. On the right visualizations of the comparisons are presented. The arrows show where they were selected from within the correlation plot. GUI: Graphical user interface, RV: Right ventricle, LV: Left ventricle, ES: End systole, ED: End diastole, Endo: Endocardial contour, Myo: Myocardial contour, Abs. ml diff: Absolute millilitre difference.

allowed for elucidating segmentation differences in order to give a detailed description of reader differences for short-axis cine images.

Backtracking CR differences in Bland–Altman plots to visualizations of segmentation differences indicated that major millilitre differences might accumulate in basal slices. Correlation plots of all metric values offered insights into qualitative reasons for RV endocardial contour disagreements. It also provided visual confirmation of the RV being difficult in the basal slices and a common cause for larger millilitre differences in CRs. The tabular metric values provided further quantitative evidence for basal slices causing the largest millilitre differences, although the apical slices are similarly difficult to segment accurately.

Furthermore, it is expected, that Lazy Luna could be helpful as a tool for CNN developers and medical experts alike. It allows for streamlining the comparison of readers in a fashion that satisfies both communities. Lazy Luna calculates accurate CRs and metric values, automatizing error-prone and time-intensive spread sheet generation. Interactive visualizations allow for understanding differences on several levels of analysis as well as suggest causal relationships between segmentation failures and CR outliers.

The Dice metric and the Hausdorff distance were taken from the surrounding literature in CNN development[9,13,14,16,32]. Two different methods were used for calculating average Dice metrics, one value concerns all images, the other concerns only images segmented by both readers. In literature it is often unclear how the Dice metric values are averaged over cases and both considerations capture relevant aspects of the segmentation task[16,33]. The metrics were extended to include the millilitre difference for the medical community, which is usually more interested in the impact of segmentation choices on volume differences.

These metrics could be arbitrarily expanded to meet other needs. Several other metrics can also be found in the surrounding literature such as the Intersection over Union[19] or the Average Surface Distance[9,17], which could be implemented accurately to apply to Shapely objects.

Pre-processing images for Lazy Luna requires manual selection due to the lack of common image-type identifiers among vendors and sequence types. Lazy Luna currently semi-automates this by presenting the user all images concerning a case in a table grouped by Dicom tags (including seriesDescription, seriesInstanceUID

	Clinical result	Mean	Std
0	LVEF [%]	− 2.6515	2.892012
1	LVEDV [ml]	− 0.1255	2.716635
2	LVESV [ml]	4.009375	4.37966
3	Dice (all slices) [%]	94.28326	2.868816
4	Dice (slices contoured by both) [%]	95.23182	1.746661
5	HD [mm]	0.652106	0.355608
6	LVM [g]	− 1.03389	4.35594
7	Dice (all slices) [%]	90.59014	6.579328
8	Dice (slices contoured by both) [%]	88.8001	6.948619
9	HD [mm]	0.849034	0.523192
10	RVEF [%]	− 0.75374	3.131688
11	RVEDV [ml]	− 2.39193	11.12953
12	RVESV [ml]	− 0.41506	6.187294
13	Dice (all slices) [%]	90.18286	5.009351
14	Dice (slices contoured by both) [%]	91.15519	3.469649
15	HD [mm]	1.628765	0.764461
16	Dice (all slices. all contours) [%]	91.90448	4.006083
17	Dice (slices contoured by both. all contours) [%]	92.15594	3.055814
18	HD (all contours) [mm]	1.082155	0.482872

Table 1. Title: Reader comparison of clinical results and segmentation metric values. Caption: Clinical result differences between readers are presented in their averages and standard deviations (in blue). They are joined with metric value averages concerning the clinical results above them (in grey). For example: the Dice values below LVEF, LVEDV, LVESV concern the LV cavity. The table presents two Dice values, one for all slices, another restricted to slices segmented by both readers. LV: Left ventricle, LVEF: Left ventricular ejection fraction, Legend: LVEDV: Left ventricular end-diastolic volume, LVESV: Left ventricular end-systolic volume, HD: Hausdorff metric, LVM: Left ventricular myocardial mass, RVEF: Right ventricular ejection fraction, RVEDV: Right ventricular end-diastolic volume, RVESV: Right ventricular end-systolic volume, Std.: Standard deviation.

	Position	Metric	LV Endocardial Contour	LV Myocardial Contour	RV Endocardial Contour
0	basal	Dice (all slices) [%]	87.99322	87.053	71.76934
1	basal	Dice (slices contoured by both) [%]	93.40139	81.29877	74.57156
2	basal	HD [mm]	1.930015	2.130997	8.128019
3	basal	Abs. ml diff. (per slice) [ml]	1.361211	0.937157	3.167208
4	midv	Dice (all slices) [%]	96.91416	91.09024	94.12743
5	midv	Dice (slices contoured by both) [%]	96.12728	89.18645	93.35773
6	midv	HD [mm]	0.835387	0.990279	2.024506
7	midv	Abs. ml diff. (per slice) [ml]	0.307792	0.421006	0.609953
8	apical	Dice (all slices) [%]	83.54174	74.24892	81.71449
9	apical	Dice (slices contoured by both) [%]	83.60426	66.52359	82.93531
10	apical	HD [mm]	0.99975	1.372196	1.579264
11	apical	Abs. ml diff. (per slice) [ml]	0.184053	0.468395	0.234856

Table 2. Title: Segmentation metric values by contour and cardiac position. Caption: The columns specify the contour type. The sections refer to different cardiac positions (defined by the first reader). The table presents two Dice values, one for all slices, another restricted to slices segmented by both readers. Legend: Midv.: Midventricular, HD: Hausdorff metric, Abs. ml diff.: Absolute millilitre difference.

and annotations by group) so that the relevant images can be selected manually. In literature, several machine-learning supported image classification methods have been experimented with to automate this task[34,35]. Pre-processing should be simplified in the future by assisting the user with automated suggestion of image types.

Training readers in CMR as well as in other fields includes curriculum-based education, simulation and competency assessment[6,36,37]. One-on-one teaching with immediate feedback is considered most effective[37]. The relevance of training has been shown to increase the quality of LV volume evaluation[6,38]. However, this type of training requires time intensive training sessions with a teacher present who explains many cases directly. That could be supported by Lazy Luna as the fast and automatic comparison of two readers may help to improve the training of trainees without direct coaching including significant time investment for manual evaluation and to bring support in the place in which additional coaching is required.

Furthermore, CNNs play an increasing role in CMR post-processing. Several confounders can complicate the automatic segmentation of images. Generalizing over different datasets can be difficult. Confounders include: different sequences such as the short-axis cine images in this paper[5,39,40], different scanners[19], different pathologies[17] (i.e. LV and RV hypertrophies) and artefacts that must be identified and excluded before automatic segmentation[1]. Lazy Luna offers functionality for the calculation of inter- and intraobserver comparisons for the assessment of segmentation accuracy.

CNNs should be compared to readers on a contour level for precise evaluation. Several CMR segmentation contests include sophisticated evaluations for segmentation quality and CRs. However, they disregard the inaccuracy caused by comparing on pixelated segmentation masks as ground truth segmentations instead of comparing contours as polygons[16,41].

CNN training procedures could integrate Lazy Luna's capabilities as part of the training procedure. By storing the annotations for the evaluation dataset in Lazy Luna's format, Dice metric values would be offered, but clinically relevant outliers of cases would also be analysed accordingly. This would enhance the evaluation by considering the interconnected nature of Dice metric values and the volumetric differences they cause.

In several guidelines it is recommended to perform evaluation based on the AHA model[1]. In the future, Lazy Luna will provide the AHA model as an intermittent analysis step, allowing for tracking of annotation differences from AHA-segments.

The classes generically keep track of images and annotations. This software backend can be extended to include other quantification techniques as well.

Limitations.

Lazy Luna is intended to be generic, however currently it is limited to short-axis cine stacks and should be shown to generalize to other cardiac structures and imaging sequences. Other outputs such as AI segmentations maps and other software vendors are to be tested in future work.

Lazy Luna is intended to be open-source in the future as to be available to and extendable by other researchers. Other image and annotation pre-processing steps (i.e. steps typically necessary for AI-contests) will be automatically addressed before source-code publication so that researchers can reproduce results on available segmentation contests.

Conclusion

The introduced software Lazy Luna enables an automatic multilevel evaluation of readers on quantitative results. In our use-case the readers showed an overall good agreement on the level of individual segmentations and clinical results. Lazy Luna allowed pinpointing origins of large millilitre difference to segmentation differences in specific cardiac structures and locations. Future developments include generalizing the software's applicability to different sequences and anatomical structures.

Data availability

The datasets analysed during the current study are not publicly available due to patient data privacy but are available from the corresponding author on reasonable request after communication with the legal department as there are special rules based on the EU law and the rules of the Berlin data officer rules. The datasets generated during this study are included in this published article and its supplementary information files.

Received: 27 January 2022; Accepted: 6 April 2022
Published online: 22 April 2022

References

1. Schulz-Menger, J. *et al.* Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update : Society for Cardiovascular Magnetic Resonance (SCMR): Board of Trustees Task Force on Standardized Post-Processing. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **22**, 19 (2020).
2. Zamorano, J. L. *et al.* 2016 ESC Position Paper on cancer treatments and cardiovascular toxicity developed under the auspices of the ESC Committee for Practice Guidelines: The Task Force for cancer treatments and cardiovascular toxicity of the European Society of Cardiology (ESC). *Eur. Heart J.* **37**, 2768–2801 (2016).
3. Zange, L. *et al.* 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. Off. J. Soc. Cardiovasc. Magn. Reson.* **21**, 12 (2019).
4. Suinesiaputra, A. *et al.* Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **17**, 63 (2015).
5. Lustig, M., Donoho, D. & Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* **58**, 1182–1195 (2007).
6. Hedström, E. *et al.* The effect of initial teaching on evaluation of left ventricular volumes by cardiovascular magnetic resonance imaging: comparison between complete and intermediate beginners and experienced observers. *BMC Med. Imaging* **17**, 33 (2017).
7. Xiong, Z. *et al.* A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Med. Image Anal.* **67**, 101832 (2021).
8. Pesapane, F., Codari, M. & Sardanelli, F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2**, 35 (2018).
9. Bai, W. *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **20**, 65 (2018).

10. Robinson, R. *et al.* Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **21**, 18 (2019).
11. Duan, J. *et al.* Automatic 3D Bi-ventricular segmentation of cardiac images by a shape-refined multi- task deep learning approach. *IEEE Trans. Med. Imaging* **38**, 2151–2164 (2019).
12. Bello, G. A. *et al.* Deep learning cardiac motion analysis for human survival prediction. *Nat. Mach. Intell.* **1**, 95–104 (2019).
13. Leiner, T. *et al.* Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **21**, 61 (2019).
14. Isensee, F. *et al.* nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. *ArXiv180910486 Cs* (2018).
15. Rajchl, M. *et al.* DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans. Med. Imaging* **36**, 674–683 (2017).
16. Bernard, O. *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. *IEEE Trans. Med. Imaging* **37**, 2514–2525 (2018).
17. Backhaus, S. J. *et al.* Fully automated quantification of biventricular volumes and function in cardiovascular magnetic resonance: applicability to clinical routine settings. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **21**, 24 (2019).
18. Sander, J., de Vos, B. D., Wolterink, J. M. & Išgum, I. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. *Med. Imaging 2019 Image Process.* (2019) https://doi.org/10.1117/12.2511699.
19. Chen, C. *et al.* Improving the generalizability of convolutional neural network-based segmentation on CMR images. *Front. Cardiovasc. Med.* **7**, 105 (2020).
20. Sander, J., de Vos, B. D. & Išgum, I. Automatic segmentation with detection of local segmentation failures in cardiac MRI. *Sci. Rep.* **10**, 21769 (2020).
21. Chen, C. *et al.* Deep learning for cardiac image segmentation: A review. *Front. Cardiovasc. Med.* **7**, 25 (2020).
22. Cardiac MRI and CT Software – Circle Cardiovascular Imaging. https://www.circlecvi.com/.
23. Mustra, M., Delac, K. & Grgic, M. Overview of the DICOM standard, in *2008 50th International Symposium ELMAR*. vol. 1, 39–44 (2008).
24. Mason, D. SU-E-T-33: Pydicom: An Open Source DICOM Library. *Med. Phys.* **38**, 3493–3493 (2011).
25. Gillies, S. & others. Shapely: manipulation and analysis of geometric objects. (2007).
26. The Shapely User Manual — Shapely 1.8.0 documentation. https://shapely.readthedocs.io/en/latest/manual.html.
27. Gillies, S. & others. Rasterio: Geospatial raster I/O for Python programmers. (2013).
28. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
29. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
30. team, T. pandas development. *pandas-dev/pandas: Pandas*. (Zenodo, 2020). https://doi.org/10.5281/zenodo.3509134.
31. Qt 5.15. https://doc.qt.io/qt-5/.
32. Valindria, V. V. *et al.* Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *ArXiv170203407 Cs* (2017).
33. Tao, Q. *et al.* Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: A multivendor, multicenter study. *Radiology* **290**, 81–88 (2019).
34 Margeta, J., Criminisi, A., Cabrera-Lozoya, R., Lee, D. C. & Ayache, N. Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition. *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.* **5**, 339–349 (2017).
35. Margeta, J. Machine learning for simplifying the use of cardiac image databases. 194.
36. Ruden, E. A., Way, D. P., Nagel, R. W., Cheek, F. & Auseon, A. J. Best practices in teaching echocardiography to cardiology fellows: a review of the evidence. *Echocardiogr. Mt. Kisco N* **33**, 1634–1641 (2016).
37. Dieden, A., Carlson, E. & Gudmundsson, P. Learning echocardiography- what are the challenges and what may favour learning? A qualitative study. *BMC Med. Educ.* **19**, 212 (2019).
38. Karamitsos, T. D., Hudsmith, L. E., Selvanayagam, J. B., Neubauer, S. & Francis, J. M. Operator induced variability in left ventricular measurements with cardiovascular magnetic resonance is improved after training. *J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson.* **9**, 777–783 (2007).
39. Vermersch, M. *et al.* Compressed sensing real-time cine imaging for assessment of ventricular function, volumes and mass in clinical practice. *Eur. Radiol.* **30**, 609–619 (2020).
40. Vincenti, G. *et al.* Compressed sensing single-breath-hold CMR for fast quantification of LV function, volumes, and mass. *JACC Cardiovasc. Imaging* **7**, 882–892 (2014).
41. Left Ventricle Full Quantification Challenge MICCAI 2019. https://lvquan19.github.io/.

Acknowledgements

We wish to thank the members of the WG CMR for input at different steps. TH receives funding from the German Research Foundation (GRK2260, BIOQIC). At time of the software's conception TH was a Master's student employee at Siemens Healthineers, Erlangen Germany.

Author contributions

All co-authors provided input to the project outline. J.W., C.A., S.D., S.L. and C.G. provided advice and support on software conception and development. M.F., A.T., F.W. and E.A. provided the sets of contours of the interobserver analysis. J.G. provided insight into clinical utility of metrics and visualizations. T.H., S.D., J.S.M., S.L., J.W. and C.G. conceptualized the software's abstract data structures. T.H. implemented the software and carried out the data analysis. All co-authors reviewed and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10464-w.

Correspondence and requests for materials should be addressed to J.S.-M.

Reprints and permissions information is available at www.nature.com/reprints.

Computer Methods and Programs in Biomedicine 238 (2023) 107615

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging

Thomas Hadler[a,b,c,*], Clemens Ammann[a,b], Jens Wetzl[d], Darian Viezzer[a,b,c], Jan Gröschel[a,b,c,e], Maximilian Fenski[a,b], Endri Abazi[a,b], Steffen Lange[g], Anja Hennemuth[c,h,i], Jeanette Schulz-Menger[a,b,c,d,e,f]

[a] *Working Group on CMR, Experimental and Clinical Research Center, a cooperation between the Max Delbrück Center for Molecular Medicine in the Helmholtz Association and Charité – Universitätsmedizin Berlin, Berlin, Germany*
[b] *Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany*
[c] *DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany*
[d] *Siemens Healthineers, Erlangen, Germany*
[e] *Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany*
[f] *Department of Cardiology and Nephrology, HELIOS Hospital Berlin-Buch, Berlin, Germany*
[g] *Department of Computer Sciences, Hochschule Darmstadt - University of Applied Sciences, Darmstadt, Germany*
[h] *Institute of Cardiovascular Computer-assisted Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany*
[i] *Fraunhofer MEVIS, Bremen, Germany*

ARTICLE INFO

Article history:
Received 14 November 2022
Revised 27 April 2023
Accepted 16 May 2023

Keywords:
Segmentation
Cardiology
Cardiac magnetic resonance imaging
Magnetic resonance imaging
Software
Quality assurance

ABSTRACT

Background and objectives: Cardiovascular Magnetic Resonance (CMR) imaging is a growing field with increasing diagnostic utility in clinical routine. Quantitative diagnostic parameters are typically calculated based on contours or points provided by readers, e.g. natural intelligences (NI) such as clinicians or researchers, and artificial intelligences (AI). As clinical applications multiply, evaluating the precision and reproducibility of quantitative parameters becomes increasingly important. Although segmentation challenges for AIs and guidelines for clinicians provide quality assessments and regulation, the methods ought to be combined and streamlined for clinical applications.

The goal of the developed software, Lazy Luna (LL), is to offer a flexible evaluation tool that is readily extendible to new sequences and scientific endeavours.

Methods: An interface was designed for LL, which allows for comparing annotated CMR images. Geometric objects ensure precise calculations of metric values and clinical results regardless of whether annotations originate from AIs or NIs. A graphical user interface (GUI) is provided to make the software available to non-programmers. The GUI allows for an interactive inspection of image datasets as well as implementing tracing procedures, which follow statistical reader differences in clinical results to their origins in individual image contours. The backend software builds on a set of meta-classes, which can be extended to new imaging sequences and clinical parameters. Following an agile development procedure with clinical feedback allows for a quick implementation of new classes, figures and tables for evaluation.

Results: Two application cases present LL's extendibility to clinical evaluation and AI development contexts. The first concerns T1 parametric mapping images segmented by two expert readers. Quantitative result differences are traced to reveal typical segmentation dissimilarities from which these differences originate. The meta-classes are extended to this new application scenario. The second applies to the open source Late Gadolinium Enhancement (LGE) quantification challenge for AI developers "Emidec", which illustrates LL's usability as open source software.

Conclusion: The presented software Lazy Luna allows for an automated multilevel comparison of readers as well as identifying qualitative reasons for statistical reader differences. The open source software LL can be extended to new application cases in the future.

[*] Corresponding author at: Grabenstr. 39, 12209, Berlin, Germany.
E-mail address: thomas.hadler@charite.de (T. Hadler).

https://doi.org/10.1016/j.cmpb.2023.107615
0169-2607/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Cardiovascular Magnetic Resonance imaging (CMR) is an ever-growing field of imaging and diagnostic techniques that are prominent in research and clinical practice [1]. These imaging techniques include cine imaging to capture motion, Late Gadolinium Enhancement (LGE) for focal scar imaging and parametric mapping techniques (i.e. T1 and T2 mapping) mainly for diffuse fibrosis and edema imaging, respectively. Focal scarring is locally concentrated scar tissue, diffuse fibrosis refers to distributed scar tissue, and edema reflects the water content in myocardial tissue.

CMR is affected by multiple confounders, which influence its processing pipeline, and require standardization and quality management. On the level of image reconstruction, standardization is approached with open-source reconstruction frameworks like Gadgetron [2]. The “Image Biomarker Standardization Initiative” aims for increased reproducibility of radiomics features in image processing [3]. Image segmentation challenges offer quality assessments of post-processing algorithms on openly available datasets [4,5]. However, these challenges often lack rigorousness due to their detachment from clinical reality. Furthermore, recommendations and consensus statements attempt to address confounders in the processing pipeline by summarizing evidence-based best practices and expert agreements [1,6–8]. Nonetheless, such statements stress the need for continuously updating recommendations for higher reproducibility of post-processing and clinical parameters in the future.

CMR diagnostic techniques rely on image processing. Although processing techniques differ in their specifics, they share key characteristics. They build on CMR images, onto which geometrical annotations are drawn. These annotations could be delineations of bloodpools, myocardium or scar tissue, or corresponding anatomical landmarks in an image sequence. Annotated images are used to calculate clinical results (e.g. amounts of fibrosis, cardiac output, etc.).

Many challenges remain for establishing and improving CMR processing methods. Manually annotating images is a laborious undertaking and training new readers is time-intensive; often including supervision by another experienced reader while analysing many training cases to reduce deviations from other experts [9,10]. Experienced readers annotate images according to the SCMR guidelines [1]. Nevertheless, significant inter- and intrareader disagreements remain due to post-processing software, site specific segmentation behaviour, difficult segmentation choices, etc. [11–14]. In order to assess precision and reproducibility of diagnostic techniques, we offer a software package capable of tracing reader differences of image annotations, over calculated clinical parameters to statistical differences (Fig. 1). The software package, called Lazy Luna (LL), allows for the comparison of exactly two readers.

In recent years, convolutional neural networks (CNN) have shown their ability to automatize image annotation tasks [5,10]. Several CNN architectures are optimized for image segmentation, such as the UNet [15] and architectural derivatives [16–18]. CNNs assign classes to image voxels. Thresholding generates segmentation masks as outputs. Although CNNs calculate clinically acceptable parameters, segmentation errors disregarding cardiac geometry remain frequent, diminishing their credibility [19,20]. The quantitative evaluation of CNN contours is habitually based on segmentation metrics like the Dice similarity coefficient (DSC) or the Hausdorff distance (HD). LL is capable of CNN evaluation and comparison as well.

The software was introduced in Hadler et al. [14]. LL was applied to a comparison of two readers on short-axis cine stacks to illustrate the workability of such comparison software. However, this did not delve into the software architecture and the interaction of its classes, or it’s capability to generalize over sequences. In order to make the software replicable and ensure it’s generalizability, this paper aims to formalize and expound on LL’s software architecture. To our knowledge, LL is the only available software package engineered towards dealing with CMR specific QA tasks, that offers a multilevel reader comparison with error-tracing as integrated software.

The aim of this paper is to design, extend and demonstrate Lazy Luna’s ability to generalize to quality assurance tasks in CMR. In order to address these tasks adequately, LL must fulfil the following criteria: LL is accessible to CNN developers and medical professionals. LL is intended as open source software and should be developed with accessible, well-known backend libraries. LL is adaptable to new imaging techniques and scientific endeavours.

2. Requirements

LL is intended as generic software capable of a multilevel reader comparison. Multilevel refers to comparing readers on the annotation level, which consists of comparing the annotations on individual images, while comparing derived clinical parameters on the patient level, and also offering a statistical level of reader comparison, i.e. to determine systematic biases (Fig. 1). LL ought to be capable of tracing differences from the image level to the reader level.

Accessibility, product independence and precision

LL needs an intuitive, generic interface for images and annotations pertaining to them so that readers can be compared, independent of vendors or reader output (polygons for human readers, image masks for CNNs). The software must offer precise calculations of quantitative results. LL is intended as open-source software and should build on available open-source components.

Adaptability and extendibility to new sequences and scientific endeavours

LL requires an understandable backend so that developers can extend the software to new sequences. LL’s core components should be broken down into classes, which allow for a generic and extendable backend.

Usability and target groups

LL should be usable by CNN developers and medical experts alike and address needs of both target groups. CNN developers and medical experts agree on the significance of calculating and displaying clinical results and their statistical evaluation. Clinicians emphasize the importance of visual inspections, whereas CNN developers focus on exact representations of contours and segmentation metrics to assess the effects of post-processing alterations. A graphical user interface (GUI) should be provided in order to allow for an automatic reader comparison with expressive visualizations and statistical analyses, independent of programming familiarity.

3. Computational methods and software architecture

3.1. Data model

In order to make LL’s backend simple to use and generic to different reader types and tasks we use folders as repositories for images and annotations pertaining to these images. The images are kept in the DICOM [21] (Digital Imaging and Communications in Medicine) format, the image annotations are stored as pickle files [22] containing Shapely [23] objects. The following subsections address these decisions separately.

DICOM is the standard for the communication and management of medical imaging information and related data [24]. The library Pydicom [25] offers a simple interface to these data. The

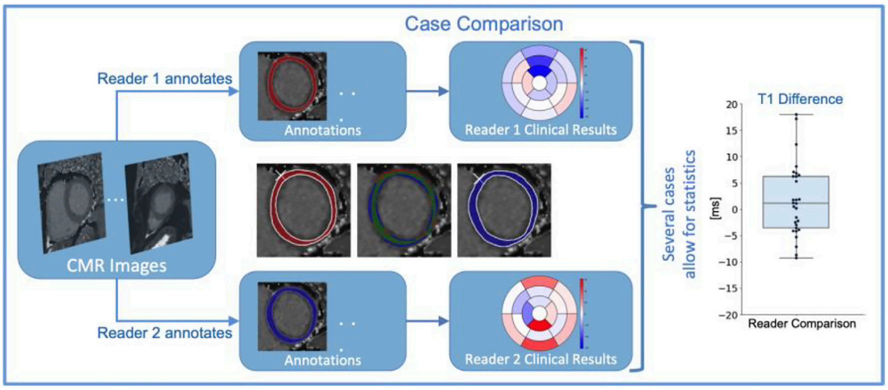


Fig. 1. A case comparison builds on two individual cases that share the same CMR images. Two readers annotate these images (reader 1 top, reader 2 bottom). The contours can be compared to each other visually (centre image) and quantitatively. Clinical results, such as the American Heart Association model segments, are calculated from a stack of contoured images. These can be compared to each other as segment value differences. When two readers analyse several cases, statistical procedures can be performed and visualized (right). Here, the quantitative parameter: T1 difference is shown. The levels of analysis are connected, allowing for the investigation of processing step influence and pipeline sensitivity.
Abbreviations: CMR: Cardiovascular magnetic resonance, AHA: American heart association.

DICOM standard allows for pooling image and image information in the same file such as the RescaleIntercept and RescaleSlope, which convert the stored pixel data to their intended output units by a linear function.

LL requires a DICOM tag for the image type (e.g. SAX T1). This is necessary for clinical practice, in which series are occasionally redone. Labelling allows for identification of the relevant series when several are available.

In order to sort images spatially or temporally, DICOM offers tags such as ImagePosition and InstanceNumber. For area and volume calculation DICOM tags provide PixelSpacing and SliceThickness.

Annotations are stored in a custom format. Every annotated image has an annotation file in pickle format. The file's basename is the referenced DICOM's unique SOPInstanceUID. The pickle file stores a python dictionary of key value pairs. Its keys are contour name strings (i.e. 'lv_myo' for LV myocardium). The keys map to geometrical Shapely objects and auxiliary information for individual contours.

3.2. Annotation processing and precision

Shapely is a Python package for manipulating and analysing geometric objects (i.e. polygons, points) [23,26]. Contours are modelled as Polygons (LV, RV endo- and epicardial contours) or MultiPolygons (papillary muscles); markers are modelled as Points (i.e. insertion points) or MultiPoints (i.e. extent points). Shapely offers precise geometrical operations including calculations, intersections, unions and Hausdorff distance (HD) calculations. Since human readers often segment on a subpixel level Shapely allows for exact geometrical calculations. In order to calculate precise values for segmentation masks (typical outputs for CNNs) the segmented pixels are outlined to produce a polygon. LL uses Rasterio's rasterize function to generate exact outlines from segmentation masks in Shapely format [27]. This permits precise calculations when comparing different reader types.

Metrics

Metrics allow for the comparison of annotations on the image level. For contours the Hausdorff Distance is calculated as the furthest distance of any point on either geometry from its nearest point on the other geometry. The Dice Similarity Coefficient is calculated as two times the intersection area divided by the sum of two Shapely geometries. The millilitres and their differences (ml Diff) are calculated using DICOM tag information on pixel height, width and slice thickness in mm and the contoured areas.

$$ml\ Diff(A,B) = (|A| - |B|) \times area\ per\ pixel\ \times\ slice\ thickness$$

$$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

$$HD(A,B) = \max\{max_{a \in cA}(\min_{b \in cB} d(\mathrm{a},\mathrm{b})),\ max_{b \in dc}(\min_{a \in cA} d(\mathrm{a},\mathrm{b}))\}$$

For points, the millimetre distance (mm Diff) is calculated. For connecting lines of specific annotation points, angular differences (angle Diff) may also be calculated:

$$mm\ Diff(p_1, p_2) = |p_1 - p_2|$$

$$angleDiff(p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2}) = arccos(u, v),$$
$$u = \frac{p_{1,1} - p_{1,2}}{|p_{1,1} - p_{1,2}|},\quad v = \frac{p_{2,1} - p_{2,2}}{|p_{2,1} - p_{2,2}|}$$

3.3. Adaptability and extendibility to new image types and scientific endeavours

3.3.1. Software class structure

LL is intended to be accessible and extendable software. LL is written in Python [28] and follows an object oriented programming paradigm. An overview of LL's software class structure and the classes' interactions are illustrated in Fig. 2. The classes are elaborated on in Hadler et al. [14].

Extending Lazy Luna classes to new image techniques

Extending LL to new imaging techniques may require alterations or extensions to several of the above backend classes. The implementation of new Clinical Results or Metrics may require alterations to underlying code, such as the Annotation class or the Category classes. Extendible classes, their main functions and

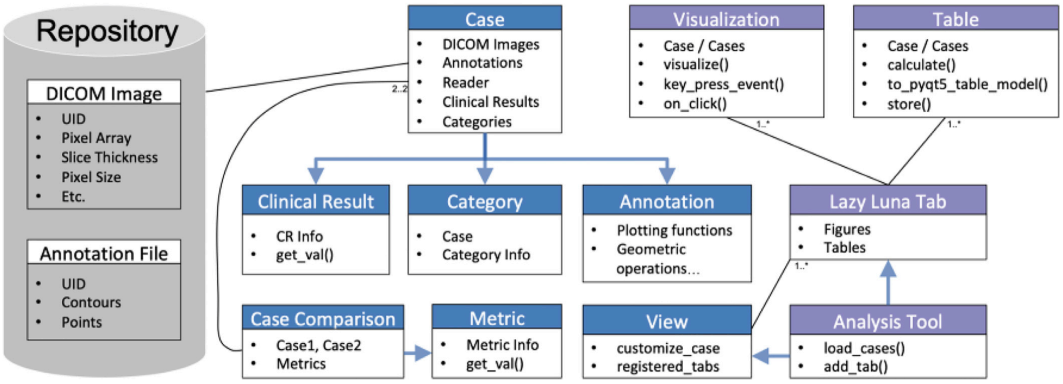


Fig. 2. Class Diagram of Lazy Luna.
Repository (grey).
The repository contains DICOM images and annotation files. The images are sorted into folders by case; the annotation files are sorted into folders by reader and case.
Backend (blue).
The class diagram depicts Lazy Luna's backend classes and how they interconnect. A Case is a container class for DICOM images and annotation files. It makes annotations accessible with an Annotation class, which offers visualization functions and geometric operations. A Case has Categories, which structure the images into slices and phases. Clinical Result classes can be attached to a Case in order to calculate CRs based on the images, annotations and categories. Case Comparisons contain two cases that reference the same images. Metric classes can be attached to a Case Comparison to calculate metric values for the case's annotations. View classes organize the other classes to address the needs of certain series. For example: a SAX Cine View can inform a case to focus on the subset of SAX Cine images and annotations, as well as attaching the respective CRs.
Frontend (purple).
Visualizations inherit their behavior from the Matplot.Figure class, allowing for user interactions. Tables inherit their behavior from the Pandas.DataFrame class. Both classes permit easy integration into PyQt5 tabs. The Analysis Tool is the GUI's central GUI. It allows for loading cases of two different readers, customizing the cases via the views and adding new tabs to the GUI.
Legend: CR: Clinical Result, GUI: graphical user interface, DICOM Image [24], Matplotlib.Figure [29], Pandas.Dataframe [30], PyQt5 [34].

extension possibilities to new endeavours are listed in Table 1. For more practical implementation details we refer to the Github repository and accompanying documentation.

Annotation: is a utility class that handles pickled dictionaries containing Shapely geometries. New geometrical calculations or visualizations would be implemented as Annotation class functions. Categories: sort images and annotations by slice and phase. Further categorizations of images, based on their temporal and spatial relationship to other images, would be implemented as Category class functions. Clinical Results: calculate clinical parameters. New imaging modalities may require new clinical parameters, implemented as individual classes. Metrics: are classes that calculate quantitative comparisons between individual annotations. New scientific endeavours may focus on new comparisons, which are implemented as individual classes. Views: organize cases to address user needs for an imaging modality. A View makes the relevant categories available (such as those that organized the fibrosis images). It connects relevant clinical results to a case. Tabs that were designed for the imaging modality are registered in Views and made available during runtime.

3.3.2. Visualizations and tables: plug-in scheme

Matplotlib Figures [29,31] and Pandas DataFrames [30,32] are used as base classes to integrate visualizations and spreadsheets into a GUI written in PyQt5 [33]. Visualizations inherit the functionality to interact with figures via events (e.g. mouse movements or key presses). Matplotlib's FigureCanvas allows for integrating LL's Visualizations into PyQt5 tabs [34]. Tables allow for exports to diverse spreadsheet formats. LL offers an interface class for PyQt5 integration via QTableViews. LL's comparison tool is the Analysis Tool; it builds on PyQt5's QTabWidget, which allows attaching QWidgets. QWidgets can be added to the Analysis Tool and embed visualizations and spreadsheets.

3.3.3. Error tracing

LL allows for connecting tables, visualizations and tabs. This enables tracing the consequences of annotation differences from the image level to the patient level. In turn, CR differences on the patient level can be traced to their contour difference origins. Alternatively, entire tabs can be opened when a specific case is to be inspected. For example, by plotting outliers in statistical plots of CRs an outlier case can be investigated for differences of individual contours.

3.3.4. Extendibility and availability

In results we demonstrate how LL was extended to fibrosis imaging as well as to focal scar imaging. For the fibrosis imaging extension, this includes devising class extensions, tables and visualizations, as well as producing GUI elements to mirror the requirements.

LL is available as open source code and published on Github. A first-use experience is described for the EMIDEC dataset, in which Late Gadolinium Enhancement was used for focal scar imaging. A use-session video is uploaded as supplementary material.

3.4. Usability

Graphical user interface

LL offers a GUI capable of creating and loading cases of two readers in order to compare them to each other.

Automated outputs

LL was developed for quick and extensive reader comparison. In order to accomplish this, each LL View offers a sequence specific storage function. This includes storing spreadsheets of metric values for annotated images, clinical results as well as outputting statistical plots.

Table 1
Class Description and Extension.
The table consists of four columns, the class name, intended class utility, main functions and a description of its extension with potential examples. The main classes are Annotation, Category, Clinical Result, Metric and View.

Class	Use Description	Functions	Extension Description
Annotation	Interface class to geometries: - Getters: for access to geometry objects - Visualizations: of geometries atop matplotlib axes - Helper functions: contain complex geometrical calculations for simple access	Getter functions: - get_contour(cont_name) - get_point(point_name) Visualization functions: - plot_contours(axis, cont_name, color) - plot_points(axis, point_name, color) - plot_face(axis, cont_name, color) - plot_cont_comparison(axis, other_anno, cont_name, colors) Helper functions: - get_contour_as_mask(cont_name)	How to: The class is extended by adding new functions Exemplary helper functions: - Point distances - Angle calculations - Bounding box determination
Category	Sorts images and annotations and offers a simple interface: - Sorting: sorts images and annotations spatially and temporally according to dicom attributes - Getters: simple interface for dicoms, images and annotations - Helper functions: contain calculations requiring sorted dicoms and annotations	Sorting function: - get_sop2depthandtime(sop_uid2filepath) Getter functions: - get_dcm(slice, phase) - get_img(slice, phase) - get_anno(slice, phase) Helper functions: - get_volume(cont_name, phase)	How to: The class is extended by adding new functions Exemplary helper functions: - Cardiac geometry descriptions (such as basal, midventricular, apical slices) - Determining phases in cardiac cycle (like end-systolic phase)
Clinical Result	A Clinical Result calculates a clinical parameters for a case Clinical parameters: calculation of values and differences between readers	Setter: - init(case): sets case, clinical parameter name, measurement unit Getters: - get_val(as_string=False) - get_val_diff(other_clinical_result, as_string=False)	How to: Lazy Luna is extended by clinical results for new imaging modalities by writing new classes Exemplary extension: - Clinical result for calculation of average myocardial voxel intensity
Metric	A Metric quantifies the difference between two annotation geometries with the support of corresponing dicoms Metric values: calculation of metric values	Setter: - init(): sets metric name, measurement unit Getter: - get_val(geo1, geo2, dcm=None, as_string=False)	How to: Lazy Luna is extended by metrics for new imaging modalities by writing new classes Exemplary extension: - Difference in number of pixels within contour type for two readers
View	A view structures cases by appending relevant categories, clinical results and tabs	Setter: - init(): sets the view name, tabs for inidividual case_comparisons and lists of case comparisons Adjust case: - initialize_case(case): calculates information requiring only one calculation - customize_case(case): connects the view's categories and clinical results Store information function: - store(case_comparisons)	How to: Extending Lazy Luna to a new imaging modality requires the implementation of a custom View class Exemplary extension: - A View for focal scar imaging

Development strategy and code maintenance

Agile development

LL's current outputs and tabs were developed in an agile development environment with clinicians in iterative feedback loops for the individual sequences. Following the implementation of the underlying backend (load images, sort them by space and acquisition time, calculate volumes from annotations), new functionalities were discussed, implemented and finally integrated or discarded in iterative loops of clinical feedback (Fig. 3).

Code maintenance

LL's code is published on Github as open source software and maintained by the first author. Github's Issues functionality allows for tracking bugs, and other forms of feedback for incremental code improvements. LL builds on several open source packages that are community maintained. As the underlying code and API of these underlying software packages evolve, LL's backend code will require adjustments that can also be effectively pursued as issues.

4. Results

The Results section will be divided into subsections corresponding to the Requirements paragraphs.

4.1. Accessibility, product independence and precision

Precision

As described in "Computational Methods and Software Architecture" and presented in Hadler et al. [14] Lazy Luna's backend offers precise calculations for annotation comparisons and calculation of clinical parameters. Polygonal contour calculations allow for geometrical accuracy, while sorting slice positions and interpolating missing slices guarantee clinical results with numerical precision. LL's backend was used for extensive comparisons of CNN architectures to a medical expert on SAX Cine images with different acquisition techniques [35].

Hardware specifications

LL was tested for 64-Bit systems of macOS Mojave 10.14.6, Windows 10 Home and Ubuntu 20.04. LL has been tested for Python 3.6 – 3.8.

Accessibility, open source software and product independence

LL builds on several open-source libraries, as described in "Computational Methods and Software Architecture". LL is available as open-source software on Github: https://github.com/thadler/LazyLuna.

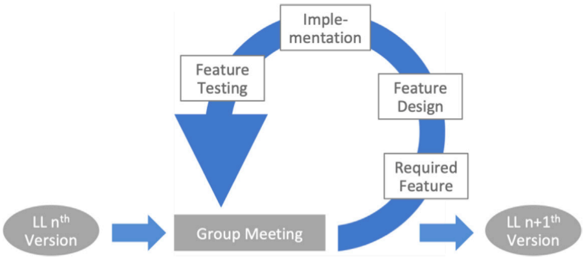


Fig. 3. Agile Software Development for Lazy Luna.
Starting at version n of LL, the incremental development consisted of team meetings with users who described concrete features they required, followed by abstractions of these to implementable features, their implementation and testing. These in turn were then kept or discarded according to the utility clinicians/AI developers saw in them in the next group meeting. After one or several such development loops a next version n+1 of LL was installed for trainee or AI evaluation.

4.2. Usability and target groups

LL offers a GUI as presented in Figs. 5,6,7,8 making it available to programmers, researchers and clinicians alike, regardless of programming skills. The software has been used in several settings to assert its utility, including trainee comparison and AI assessment, as follows. The software's graphical user interface has been used by clinicians in the working group for comparison between trainees and expert readers to provide quality assurance and standardization of contouring techniques in research and clinical routine. LL was also used for an extensive comparison between two readers on an in-house dataset [14]. LL's backend was used for extensive comparisons of AIs with different backend CNN architectures to a medical expert on SAX Cine images with different acquisition techniques [35].

The rest of the Results section will focus on the requirement of LL's adaptability and extendibility by presenting different steps of an extension to fibrosis and edema imaging (T1 & T2 mapping) as well as focal scar imaging (LGE).

4.3. Adaptability and extendibility to new sequences and scientific endeavours

LL's class structure is extendable to new sequences. By inheriting from the base classes for specific purposes LL can be extended. We demonstrate such extensions, first, for fibrosis imaging, second for focal scar imaging.

4.3.1. Extending Lazy Luna to fibrosis imaging

The adjustments largely mirror typical requirements of T1 mapping [36]. This type of parametric technique often entails locating quantitative difference within the cardiac geometry (Fig. 4).

LL's extension will reflect this in the extended classes below:

- Annotation: T1 mapping requires contours of the myocardium and an insertion point, which demarcates where the right ventricle connects to the left ventricular (LV) myocardium. The insertion point and the centroid of the LV endocardium divide the LV myocardium into segments by degrees. The Annotation class allows for calculating arbitrary numbers of segments
- Category: Images are typically acquired as stacks that span the length of the heart from base to apex. Sorting images allows for locating abnormalities/pathologies and defining segments according to the American Heart Association model [37]. The Category class must be extended to sort the images according to cardiac location
- Clinical Result: The average of T1 values in ms for annotated images
- Metrics: The Dice Similarity Coefficient and the Hausdorff metric are transferred from the SAX Cine applications. The myocardium's average pixel values and their differences are added for T1 mapping, as well as the insertion point to LV centroid angle differences between both readers
- Visualizations: One figure visualizes segmentation and insertion point differences between readers by plotting a coloured comparison of differences (Fig. 6). A second figure produces the AHA model according to annotations of the respective readers. A third figure presents histograms for different segment averages, depending on the insertion point positioning
- Tables: LL offers a table of T1 value averages and average differences for myocardial segments for both readers
- Tabs: A tab for an average AHA model for all cases and their differences is offered (Fig. 7). Another tab for individual cases allows the inspection of the effect of segmentation differences and insertion point differences on the T1 averages of arbitrary numbers of segments (Fig. 8).

Plug-in concept for tables and visualizations

A case overview tab was implemented. It shows a table of the case's clinical results and images with annotations plotted on top. The user can click through the image stack with up/down arrow keys. The Table uses LL's Clinical Result classes. The visualization builds on a T1 mapping Category for slice sorting/accessing images by slice, image correction by rescaling the image values according to DICOM attributes and the Annotation class' support functions for contour plotting (Fig. 5).

Metrics, qualitative comparisons and error tracing

LL admits for tracing reader differences from the patient level to the image level by connecting statistical analysis tabs to case specific tabs. Statistical visualizations, which plot cases as points (e.g. Bland-Altman, correlation plots, etc.), can interactively open tabs with additional information pertaining to the case (Fig. 6).

AHA model and error tracing

The 16-segment AHA model is a popular geometrical abstraction of the heart that is used in the clinical environment to present local average T1 mapping values. It sorts the image stack into basal, midventricular and apical slices, and divides individual slices into six (for basal, midventricular) and four (for apical) segments. This allows for tracing the global value of T1 parametric values from the patient level to the position of T1 value outliers. In order to consider image segments the Annotation class was extended to allow for their calculation. This requires the myocardial contour and the reference point.

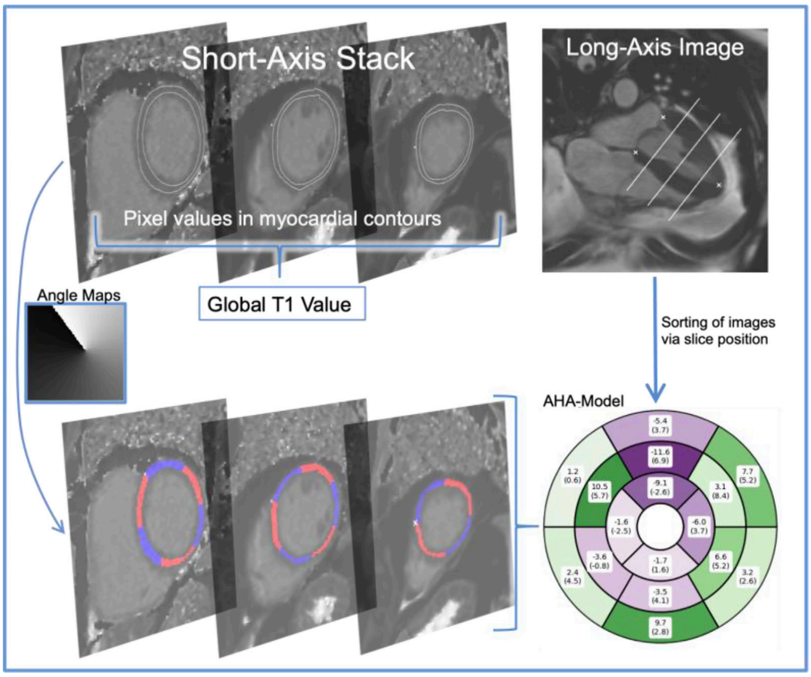
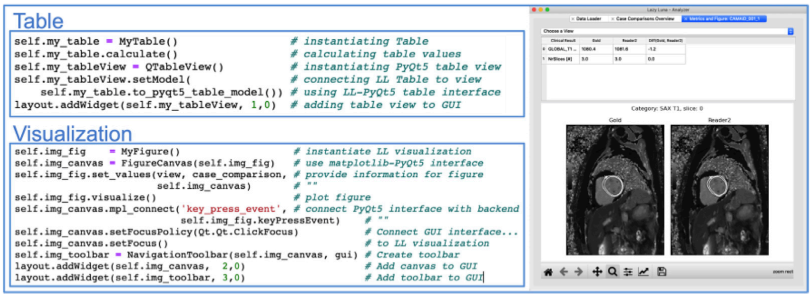


Fig. 4. Pipeline on calculating the AHA model from DICOM images with Annotations.
At the top left there are T1 parametric mapping images with contours and insertion points pertaining to them. The myocardium is divided into segments (red, blue) via angle maps, which are derived from the annotations (lower left). The LVM mask is generated from contours by selecting all pixels whose centre is within the polygon or by Bresenham's line algorithm. The individual images are assigned to a basal, midventricular or apical location depending on the their spatial relationship to the extent and apical points in a long-axis view of the heart (top right). On the bottom right the AHA model is calculated by assigning sorting the segments into their respective bins and calculating the average. The rings correspond to the basal, midventricular and apical positions (outside to inner).
Legend: AHA: American Heart Association, LVM: Left ventricular myocardium.



Table

```
self.my_table = MyTable()                         # instantiating Table
self.my_table.calculate()                         # calculating table values
self.my_tableView = QTableView()                  # instantiating PyQt5 table view
self.my_tableView.setModel(                       # connecting LL Table to view
    self.my_table.to_pyqt5_table_model()) # using LL-PyQt5 table interface
layout.addWidget(self.my_tableView, 1,0)  # adding table view to GUI
```

Visualization

```
self.img_fig    = MyFigure()                      # instantiate LL visualization
self.img_canvas = FigureCanvas(self.img_fig)      # use matplotlib-PyQt5 interface
self.img_fig.set_values(view, case_comparison,    # provide information for figure
                        self.img_canvas)          # ""
self.img_fig.visualize()                          # plot figure
self.img_canvas.mpl_connect('key_press_event',    # connect PyQt5 interface with backend
                            self.img_fig.keyPressEvent)    # ""
self.img_canvas.setFocusPolicy(Qt.Qt.ClickFocus)           # Connect GUI interface...
self.img_canvas.setFocus()                                 # to LL visualization
self.img_toolbar = NavigationToolbar(self.img_canvas, gui) # Create toolbar
layout.addWidget(self.img_canvas,  2,0)                    # Add canvas to GUI
layout.addWidget(self.img_toolbar, 3,0)                    # Add toolbar to GUI
```

Lazy Luna - Analyzer
× Data Loader
× Case Comparisons Overview
× Metrics and Figure: CAMAD_001_1
Choose a View

	Clinical Result	Gold	Reader2	Diff(Gold, Reader2)
0	GLOBAL_T1 ...	1080.4	1081.6	-1.2
1	NrSlices [#]	3.0	3.0	0.0

Fig. 5. Adding Tables and Visualizations to LL Analysis Tool.
On the left code samples are exemplified; on the right the resulting tab of the LL Analysis Tool is presented. On the upper left, code for the integration of an LL Table into the Analysis Tool is shown. After implementing an LL Table it can be instantiated and added to a tab by using QTableViews as an interface. On the lower left, code for the integration of an LL Visualization into the Analysis Tool is presented. After writing an LL Visualization, it can be instantiated and added to a tab by using the FigureCanvas class as an interface to PyQt5. The GUI can be connected to the visualization's key press events with the function mpl_connect. On the right, the Analysis tool is presented with the tab containing the table and visualization. Further code examples can be viewed in github online.
Legend: LL: Lazy Luna, QTableView [34], FigureCanvas [29], GUI: graphical user interface, Global_T1: Average of T1 values inside myocardium for all slices in image stack.

In order to compare two readers at assessing the reproducibility of the AHA segments, LL provides AHA calculations and visualizations for individual cases as well as reader differences by segment. Analogously, for patient cohorts, averaged AHA models and averages of the differences between both readers' segment values are presentable (Fig. 7).

Number of myocardial segments and insertion point and error tracing

As above, calculating average T1 values in myocardial segments of images is central to post-processing of fibrosis images to locate abnormal ranges. Average segment value differences between readers can be caused by contour deviations or different insertion

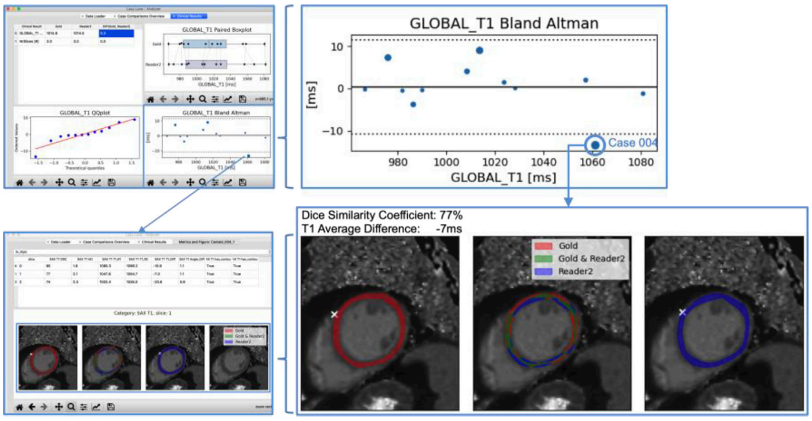


Fig. 6. Error Tracing and Differences.
On the left two, Lazy Luna tabs are presented. On the right, outlined GUI parts are expanded. The upper left tab provides a table of average CR values and statistical plots of the analysed cases, including a paired boxplot, a QQ-plot and a Bland-Altman plot of global T1 values. The Bland-Altman plot is magnified with the circled case being the largest outlier (upper right). By clicking the case the lower tab was opened for an in-depth inspection (lower left). A table of metric values per slice is presented on the top, and a visualization of reader contours and their agreement/disagreement below (gold reader red, reader 2 blue, reader agreement green). The reader differences, as quantified by the DSC and the T1 Average Difference are caused by these contour differences, revealing the origin of the global T1 value difference of the outlier case.
Legend: LL: Lazy Luna, QTableView[32], FigureCanvas[30], GUI: graphical user interface, Global_T1: Average of T1 values inside myocardium for all slices in image stack.

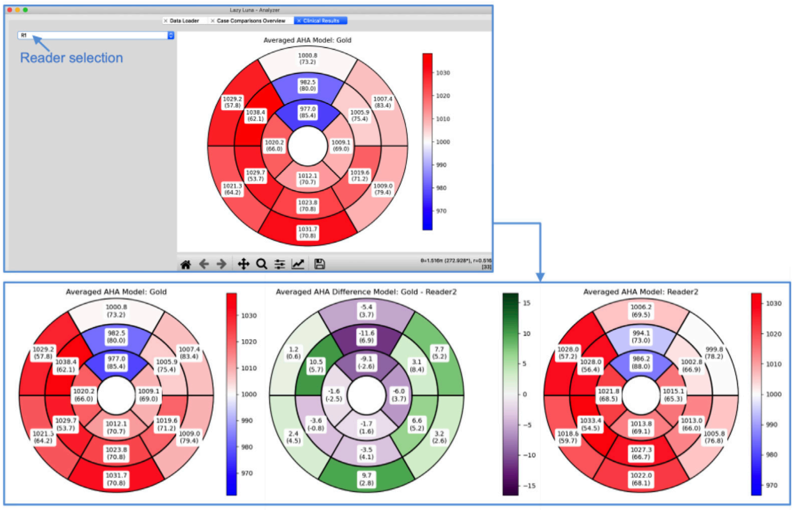
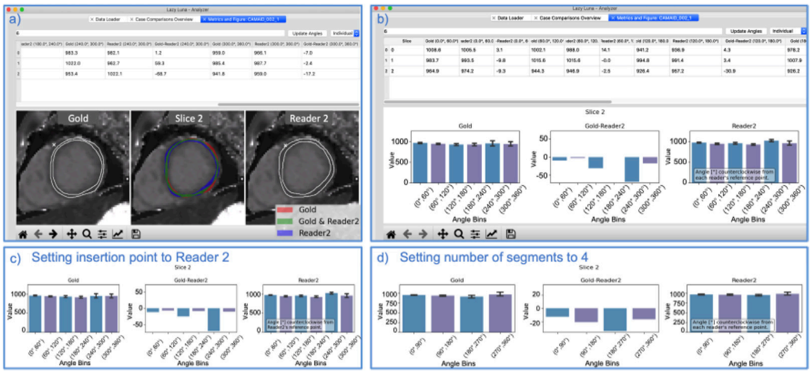


Fig. 7. AHA model.
On the top the Average AHA Model tab for several patients shows the average (and the standard deviation) of segment values of a reader for all provided cases. The reader was selected on the upper left of the GUI. Below three figures generated from this tab are presented (from left to right): First, the average AHA model for all cases for the first reader is presented. Second, the average of differences between the first and second reader is visualized. Third, the average AHA model for all cases for the second reader.
Legend: AHA: American heart association, GUI: graphical user interface.

points, which position the segments along the myocardium. Varying numbers of segments may affect the value averages since the number of pixels per segment decreases when the number of segments grows. LL's backend classes offer functions to assess these segment-level confounders. LL's GUI allows the user to examine the effects of these confounders (Fig. 8).

4.3.2. *Extending Lazy Luna to focal scar imaging and CNN outputs*

An LL jupyter notebook was used to convert Emidec's image dataset from Nifti format to DICOM format and the segmentation masks to LL's annotation format. The Nifti format consists of a header with the image's meta information and the image data. The header contains relevant information such as voxel width, height,


Lazy Luna - Analyzer
a)
× Data Loader
× Case Comparisons Overview
× Metrics and Figure: CAMAID_002_1
6
Update Angles
Individual

	ader2 (180.0°, 240.0°)	Gold (240.0°, 300.0°)	Reader2 (240.0°, 300.0°)	Gold-Reader2 (240.0°, 300.0°)	Gold (300.0°, 360.0°)	Reader2 (300.0°, 360.0°)	Gold-Reader2 (300.0°, 360.0°)
0		983.3	982.1	1.2	959.0	966.1	-7.0
1		1022.0	962.7	59.3	985.4	987.7	-2.4
2		953.4	1022.1	-68.7	941.8	959.0	-17.2

	Slice	Gold (0.0°, 60.0°)	ader2 (0.0°, 60.0	-Reader2 (0.0°, 6	old (60.0°, 120.0°	ader2 (60.0°, 120.	Reader2 (60.0°, 1	old (120.0°, 180.0	Reader2 (120.0°, 180.0°)	Gold-Reader2 (120.0°, 180.0°)	Gold (18
0	0	1008.6	1005.5	3.1	1002.1	988.0	14.1	941.2	936.9	4.3	978.2
1	1	983.7	993.5	-9.8	1015.6	1015.6	-0.0	994.8	991.4	3.4	1007.9
2	2	964.9	974.2	-9.3	944.3	946.9	-2.5	926.4	957.2	-30.9	926.2

Fig. 8. Tracing Segment Differences.
The Segments and Insertion Point tab allows to switch between two different visualizations (a). The first line of the tab consists of the (left to right) selecting the number of segments for the myocardium, an update button for the table and figure below, and a selection option for the insertion point (first reader's insertion point, the second reader's, or each individual reader's point). The second row is a table of segment values; three consecutive columns concern the first reader's segment average, the second reader's and their difference. The main figure (row three) shows contours and insertion points for both readers, which are plotted on the left and right image, respectively. The centre image presents a contour comparison of both readers. In b) the figure of tab a) was switched (via shift key) to histograms of the segments' average values: The first and third histograms concern the readers' average segment values; the middle histogram represents the average segment value differences. In c) histograms were created according to the second reader's insertion point for both readers, which led to minor histogram value deviations. In d) a histogram was calculated with four instead of six histograms, which "averaged away" 30ms of reader differences.

and depth. The image data has a shape, which provides the image sizes, the number of phases and slices. With this information a Dicom file is constructed. The annotation files were generated from the Nifti file masks, stored in the same format as the images. The voxel segmentations in Nifti format were outlined as polygons and stored in LL annotation format as pickle files, mapping contour names to shapely geometries. The exact code is published in the repository as a Jupyter Notebook and may guide new users to extending LL to new data formats.

The annotation class required no changes to generalize to focal scar imaging. Several new clinical results were implemented as classes, including the LV volume, LVM volume and mass, scar volume, mass and fraction, excluded volume as mass as well as no reflow volume. LL also required a new View class, the SAX_LGE_View to refocus the cases on the images and clinical parameters.

A UNet predicted segmentation masks for myocardium, scar and no reflow tissue. We used this under-trained network to predict the Emidec segmentations for the training set. We used LL to produce the reader comparisons in Fig. 9 and offer an investigative video in supplementary material.

5. Discussion

Lazy Luna (LL) is a software package that offers a multilevel comparison of readers on different CMR techniques. It is available to clinicians and programmers alike. LL's extendibility to new image types and scientific endeavours was illustrated by presenting a step-by-step extension of LL to fibrosis imaging. In results this allowed for an illustrative reader comparison on fibrosis imaging cases. LL has been used in other settings to assert its functionality and utility. The GUI has successfully been used for comparison between trainees and expert readers to provide quality assurance and standardization of contouring techniques in research and clinical routine. LL was also used for an extensive comparison between two readers on an in-house dataset [14].

LL provides error tracing by connecting different analysis levels. We demonstrated this on an AHA model difference for individual case comparison and an average of AHA model differences for two readers. By doing this, the relevance of contouring differences can be traced from individual images and segments to reader trends, while offering insights into how segmentation difficulty and cardiac geometry interact.

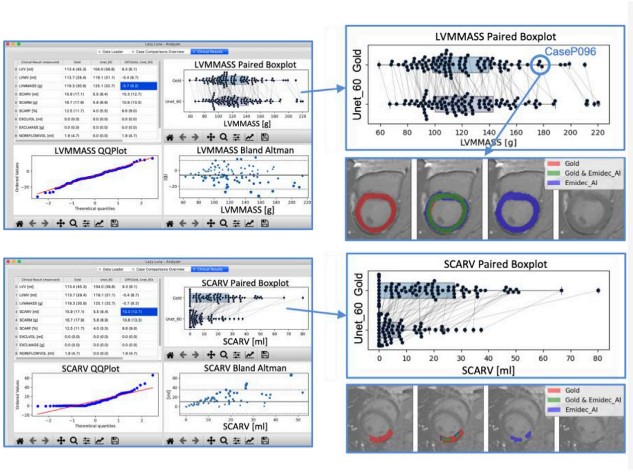
Increasingly, AIs are being applied to multiple sequences to quantify several cardiac parameters simultaneously: such as SAX Cine imaging, LAX Cine-, fibrosis-, edema- and scar imaging [38,39]. As this becomes more prominent, full sequence comparisons as offered by LL should become more typical. LL is provided as open-source software. Its usability and general concept can be verified quickly with the EMIDEC dataset on Github. LL should allow AI developers to evaluate the quality of their algorithms on several levels of analysis simultaneously, satisfying the interest of CNN developers while addressing the clinical relevance of differences.

In recent years, CNN developers have argued that their algorithms are within the range of interobserver reproducibility and thus also in typical variability of clinical routine parameters [5,16]. The equivalence of different contouring software has been assessed by testing that confidence intervals were within defined tolerance ranges [11]. Such ranges are necessarily parameter specific as different confounders have different effect-sizes. Assessing and defining tolerable biases between CNNs and clinicians as well as limiting the proportion of cases, which lie outside of such tolerance ranges, may be the topic of another work. LL should be extended to test for, and visualize, reader equivalence according to well-defined criteria.

Likewise, the training and education of readers is a time-intensive task, based on curriculums and proficiency assessments [10,40]. One-on-one teacher-student training is deemed most advantageous, but is also most resource absorbing [9,10]. Training and education has been shown to improve LV volume reproducibility [9]. LL could be used in training settings to automate difference assessment as well as offering illustrations of annotation differences between teacher and student.

Outlook

LL can analyse and characterise reader differences in order to standardize contouring methods for more reproducible clinical results. AIs are increasingly relevant for quantifying CMR images. LL has been used to investigate different CNN architectures' (UNet



Lazy Luna - Analyser

	Clinical Result (measured)	Gold	Unet_60	Diff(Gold, Unet_60)
0	LVV [ml]	113.4 (45.3)	104.0 (39.8)	9.3 (8.1)
1	LVMV [ml]	113.7 (29.4)	119.1 (31.1)	-5.4 (8.7)
2	LVMMASS [g]	119.3 (30.9)	125.1 (32.7)	-5.7 (9.2)
3	SCARV [ml]	15.9 (17.1)	5.5 (8.4)	10.3 (12.7)
4	SCARM [g]	16.7 (17.9)	5.8 (8.9)	10.8 (13.3)
5	SCARF [%]	12.5 (11.7)	4.0 (5.5)	8.6 (9.0)
6	EXCLVOL [ml]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
7	EXCLMASS [g]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
8	NOREFLOWVOL [ml]	1.9 (4.7)	0.0 (0.0)	1.9 (4.7)

Fig. 9. Emidec CNN Analysis.
The top three sub-figures compare the UNet's LVM estimation to the gold standard's. The Clinical Results tab (top left) offers an overview of all clinical result averages of the gold standard and the UNet_60 as well as their differences. The focus lies on the paired boxplot enlarged on the upper right. Here, the two readers (Gold top, Unet_60 bottom) are presented as boxplots with the case's LVM values plotted as dots on top. Lines connect the case dots in order to visualize the case-specific reader differences. Below the reader contours of CaseP096 were presented, showing the constant overestimation of the epicardial contour, providing a qualitative clue to the LVM difference. The bottom three sub-figures provide an analogous analysis concerning the UNet's SCARV estimation. The paired boxplot reveals that the UNet consistently underestimates the scar volume drastically. The lower plot reveals that the UNet has not yet learned to estimate the full scar but only fragments of it.
Legend: LVMMASS: left ventricular mass, SCARV: scar volume, ml: millilitre, g: gram.

[15], FCN [41], Dilated UNet [42], MultiResUnet [43]) performances against expert readers [35]. Several working groups have enhanced CNN performance by integrating cardiac geometry assumptions or plausibility checks of the heart's blood pool volumes over the cardiac cycle into the overall segmentation pipeline [44–46]. Whether AI performance improves by integrating cardiac geometry information should be investigated.

We based LL on a class diagram for explainability. The general idea is intuitive: Cases contain images and annotations, categories manage the images and annotations, and the images and annotations can be combined to calculate clinical results per case. When two cases reference the same images the annotations can be compared to each other on the image level, and the clinical results on patient level. This applies to CMR as we could generalize to different sequences in Results. Furthermore, this principle applies to any conceivable 2D imaging modality, such as Neuro MRI, CT scans or echocardiography.

Releasing the software as open source code should allow for more feedback and a sharing of maintenance costs for LL. This should provide incentives to improve and adjust the software to a variety of needs.

Limitations

LL requires user intervention to recognize image types (such as SAX Cine or T1 Mapping images). Currently, this recognition is semi-automated by a DICOM tag focussed image-type suggestion. However, users have described it as tedious and the task ought to be automatized in future work. The current version of LL allows for the comparison between exactly two readers (regardless of them being human or AI). Comparing multiple readers to another reader (often the gold standard reader) would be a useful extension to LL. This study is limited to the analysis of LL's software architecture. Case studies that would illustrate the utility of LL in QA scenarios are out of the scope of this work's scope.

6. Conclusion

The presented software Lazy Luna allows for a multilevel reader comparison on several sequences typical in the clinical domain, while remaining flexible and extendible to new scientific undertakings. Extending LL to T1 parametric mapping demonstrated the software's flexibility. LL will enhance reader comparisons by merging AI analysis with clinical analysis and contribute to standardization and reproducibility in clinical routine.

Author contributions

All co-authors provided input to the project. CA, JW, DV, SL, AH and JSM provided advice and support on software development. MF and EA provided the contours of the presented cases. JG and JSM provided feedback on the software's utility in practice. TH implemented the software and carried out the data analysis. All coauthors reviewed and approved the final manuscript.

Ethics declarations

The local ethics committee of Charité Medical University Berlin gave ethics approval for the original study (approval number EA1/323/15). All patients gave their written informed consent before participating in the study.

Data availability

We have all source images and workspaces. They are saved at a central server of the Charité as well as locally on a working group specific server. We could make access possible following dedicated request after communication with the legal department as there are special rules based on the EU law and the rules of the Berlin data officer rules. All the data generated during this study are included in this published article and its supplementary information files.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements including Declarations

We wish to thank the members of the WG CMR for feedback and input. TH receives funding from the German Research Foundation (GRK2260, BIOQIC). At time of the software's conception TH was a Master's student employee at Siemens Healthineers, Erlangen Germany.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107615.

References

[1] J. Schulz-Menger, et al., Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update : Society for Cardiovascular Magnetic Resonance (SCMR): Board of Trustees Task Force on Standardized Post-Processing, J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson. 22 (2020) 19.
[2] M.S. Hansen, T.S. Sørensen, Gadgetron: An open source framework for medical image reconstruction: Gadgetron, Magn. Reson. Med. 69 (2013) 1768–1776.
[3] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image biomarker standardisation initiative, Radiology 295 (2020) 328–338.
[4] Left Ventricle Full Quantification Challenge MICCAI 2019. https://lvquan19.github.io/.
[5] O. Bernard, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (2018) 2514–2525.
[6] P. Haaf, et al., Cardiac T1 Mapping and Extracellular Volume (ECV) in clinical practice: a comprehensive review, J. Cardiovasc. Magn. Reson. 18 (2017) 89.
[7] D.R. Messroghli, et al., Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: a consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI), J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson. 19 (2017) 75.
[8] on behalf of SCMR Clinical Trial Writing Group et al. Society for Cardiovascular Magnetic Resonance (SCMR) expert consensus for CMR imaging endpoints in clinical research: part I - analytical validation and clinical qualification. *J. Cardiovasc. Magn. Reson.* 20, 67 (2018).
[9] T.D. Karamitsos, L.E. Hudsmith, J.B. Selvanayagam, S. Neubauer, J.M. Francis, Operator induced variability in left ventricular measurements with cardiovascular magnetic resonance is improved after training, J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson. 9 (2007) 777–783.
[10] E. Hedström, et al., The effect of initial teaching on evaluation of left ventricular volumes by cardiovascular magnetic resonance imaging: comparison between complete and intermediate beginners and experienced observers, BMC Med. Imaging 17 (2017) 33.
[11] L. Zange, et al., 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. Off. J. Soc. Cardiovasc. Magn. Reson. 21 (2019) 12.
[12] S. Marchesseau, J.X.M. Ho, J.J. Totman, Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: a reproducibility study, Eur. J. Radiol. Open 3 (2016) 60–66.
[13] J. Mullally, et al., Marked variability in published CMR criteria for left ventricular basal slice selection - impact of methodological discrepancies on LV mass quantification, J. Cardiovasc. Magn. Reson. 15 (2013) P101.
[14] T. Hadler, et al., Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging, Sci. Rep. 12 (2022) 6629.
[15] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. *ArXiv150504597 Cs* (2015).
[16] W. Bai, et al., Automated cardiovascular magnetic resonance image analysis with fully convolutional networks, J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson. 20 (2018) 65.
[17] Isensee, F. et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. *ArXiv180910486 Cs* (2018).
[18] J. Duan, et al., Automatic 3D Bi-ventricular segmentation of cardiac images by a shape-refined multi- task deep learning approach, IEEE Trans. Med. Imaging 38 (2019) 2151–2164.
[19] Shwartzman, O., Gazit, H., Shelef, I. & Riklin-Raviv, T. The Worrisome Impact of an Inter-rater Bias on Neural Network Training. *ArXiv190611872 Cs Eess* (2020).
[20] J. Sander, B.D. De Vos, J.M. Wolterink, I. Išgum, Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI, Med. Imaging 2019 Image Process 44 (2019), doi:10.1117/12.2511699.
[21] DICOM. *DICOM* https://www.dicomstandard.org.
[22] G. Van Rossum, The Python Library Reference, release 3.8.2, Python Software Foundation, 2020.
[23] Gillies, S. & others. Shapely: manipulation and analysis of geometric objects. (2007).
[24] M. Mustra, K. Delac, M. Grgic, Overview of the DICOM standard, in: 2008 50th International Symposium ELMAR, 1, 2008, pp. 39–44.
[25] D. Mason, SU-E-T-33: Pydicom: an Open Source DICOM Library, Med. Phys. 38 (2011) 3493.
[26] The Shapely User Manual — Shapely 1.8.0 documentation. https://shapely.readthedocs.io/en/latest/manual.html.
[27] Gillies, S. & others. Rasterio: geospatial raster I/O for Python programmers. (2013).
[28] G. Van Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, 2009.
[29] matplotlib.figure.Figure — Matplotlib 3.3.4 documentation. https://matplotlib.org/3.3.4/api/_as_gen/matplotlib.figure.Figure.html.
[30] pandas.DataFrame — pandas 1.4.1 documentation. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html.
[31] J.D. Hunter, Matplotlib: A 2D Graphics Environment, Comput. Sci. Eng. 9 (2007) 90–95.
[32] team, T. pandas development. pandas-dev/pandas: Pandas. (2020) doi:10.5281/zenodo.3509134.
[33] Qt 5.15. https://doc.qt.io/qt-5/.
[34] PyQt - QTab Widget. https://www.tutorialspoint.com/pyqt/pyqt_qtabwidget.htm.
[35] Hadler, T., Amman, C., Gröschel, J. & Schulz-Menger, J. Multilevel comparison of neural networks for ventricular function quantification in CMR accelerated by compressed sensing. *ISMRM - Int. Soc. Magn. Reson. Med.*
[36] D.R. Messroghli, et al., Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: a consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI), J. Cardiovasc. Magn. Reson. 19 (2017) 75.
[37] Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart. 4.
[38] Multi-sequence myocardium segmentation with cross-constrained shape and neural network-based initialization, Comput. Med. Imaging Graph 71 (2019) 49–57.
[39] A Chartsias, et al., Disentangle, align and fuse for multimodal and semi–supervised image segmentation, IEEE Trans. Med. Imaging 40 (2021) 781–792.
[40] E.A. Ruden, D.P. Way, R.W. Nagel, F. Cheek, A.J. Auseon, Best practices in teaching echocardiography to cardiology fellows: a review of the evidence, Echocardiogr. Mt. Kisco N 33 (2016) 1634–1641.
[41] Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *ArXiv14114038 Cs* (2015).
[42] Wang, S. et al. U-Net Using Stacked Dilated Convolutions for Medical Image Segmentation. 8.
[43] N. Ibtehaz, M.S. Rahman, MultiResUNet : rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[44] R. Robinson, et al., Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study, J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Reson. 21 (2019) 18.
[45] B. Ruijsink, et al., Quality-aware semi-supervised learning for CMR segmentation, Stat. Atlases Comput. Models Heart STACOM Workshop (2020) 97–107 2020.
[46] Chen, C. et al. Learning Shape Priors for Robust Cardiac MR Segmentation from Multi-view Images. in vol. 11765 523–531 (2019).

Curriculum Vitae

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publication List

Original Research Papers:

1. **Hadler T**, Wetzl J, Lange S, Geppert C, Fenski M, Abazi E, et al. Introduction of Lazy Luna an automatic software-driven multilevel comparison of ventricular function quantification in cardiovascular magnetic resonance imaging. Sci Rep. 2022 Dec;12(1):6629.
   IF: Scientific Reports 4.996 (2021)
2. **Hadler T**, Ammann C, Wetzl J, Viezzer D, Gröschel J, Fenski M, et al. Lazy Luna: Extendible software for multilevel reader comparison in cardiovascular magnetic resonance imaging. Computer Methods and Programs in Biomedicine. 2023 Aug;238:107615.
   IF: Computer Methods and Programs in Biomedicine 7.027 (2021)
3. Ammann C*, **Hadler T***, Gröschel J, Kolbitsch C, Schulz-Menger J. Multilevel comparison of deep learning models for function quantification in cardiovascular magnetic resonance: On the redundancy of architectural variations. Frontiers in Cardiovascular Medicine [Internet]. 2023 [cited 2023 Jun 5];10. Available from: https://www.frontiersin.org/articles/10.3389/fcvm.2023.1118499
   – Shared first authorship between Clemens Ammann and Thomas Hadler
   IF: Frontiers in Cardiovascular Medicine 5.846 (2021)
4. Viezzer D, **Hadler T**, Ammann C, Blaszczyk E, Fenski M, Grandy TH, et al. Introduction of a cascaded segmentation pipeline for parametric T1 mapping in cardiovascular magnetic resonance to improve segmentation performance. Sci Rep. 2023 Feb 6;13(1):2103.
   IF: Scientific Reports 4.996 (2021)
5. Gröschel J, Trauzeddel RF, Müller M, Von Knobelsdorff-Brenkenhoff F, Viezzer D, **Hadler T**, et al. Multi-site comparison of parametric T1 and T2 mapping: healthy travelling volunteers in the Berlin research network for cardiovascular magnetic resonance (BER-CMR). J Cardiovasc Magn Reson. 2023 Aug 14;25(1):47.
   Journal of Cardiovascular Magnetic Resonance (6.903)
6. Gröschel J, Kuhnt J, Viezzer D, **Hadler T**, Hormes S, Barckow P, et al. Comparison of manual and artificial intelligence based quantification of myocardial strain by feature tracking—a cardiovascular MR study in health and disease. Eur Radiol [Internet].

2023 Aug 18 [cited 2023 Oct 8]; Available from: https://link.springer.com/10.1007/s00330-023-10127-y

IF: European Radiology 7.034 (2021)

Presentations and Posters (Scientific Congresses)

1. Ammann C, **Hadler T**, Gröschel J, Schulz-Menger J. Multilevel Evaluation of Four Convolutional Neural Network Architectures for Ventricular Function Quantification from Short-Axis Cine Images. SCMR 25th Annual Scientific Sessions; Online (due to Covid).
2. Viezzer D, **Hadler T**, Blaszczyk E, Fenski M, Gröschel J, Lange S, et al. Improving U-Net based segmentation in cardiac parametrical T1 mapping by incorporating bounding box information. ISMRS-ESMRMB ISMRT 31st Annual Meeting; 2022 May; London, Great Britain.
3. **Hadler T**, Amman C, Gröschel J, Schulz-Menger J. Multilevel Comparison of Neural Networks for Ventricular Function Quantification in CMR accelerated by Compressed Sensing.
4. **Hadler T**, Gröschel J, Lange S, Schulz-Menger J. Dropunet for Segmentation Quality Estimation: Towards Reflective AI for Quantification in Short-Axis Cine Images. SCMR 26th Annual Scientific Sessions; 2023 Jan; San Diego, California, USA.

Acknowledgments

Luck, often underestimated, can wield a profound influence on the trajectory of one's life and career. A kind of "cosmic" luck tosses us into the world at a specific point in time and space, provides us with a family and creates us with luck of a genetic draw. During our education luck may lead us to more or less suitable supervisors, and occasionally even to benevolent mentors. I cannot express enough gratitude for the luck I've had in life, but I shall thank the people who have mentored me as well as those who have loved me.

My PhD Thesis would not have been possible without my first supervisor, Prof. Dr. med. Jeanette Schulz-Menger, who's ideas, criticisms, and support guided all presented projects and scientific endeavours. Her unique mixture of benevolence and determination has been an invaluable moral compass to me work- and life-wise. I hope to enjoy her support for years to come. I wish thank Prof. Dr. Anja Hennemuth and Dr. Christoph Kolbitsch who both supervised my publications with scientific rigor, and invested invaluable time into polishing argumentations and presentation methods. Lastly, I would be amiss not to thank Prof. Dr. Steffen Lange who is another unique personality mixture of down-to-earthness and brilliance. He is a mentor who went above and beyond, and guided me through two theses emotionally and scientifically.

Outside the realm of supervision, I share my days with colleagues and friends from the Working Group Cardiac MRI with whom I spend hours of discussions that feed new scientific ideas and projects. Further, I wish to thank the group's doctors and technicians who perform the pristine and foundational work on which all data science builds.

Last, but not least, I am deeply grateful for my personal relationships. My parents are loyal supporters of all my undertakings, in better and worse times – a feat for which I can never express enough appreciation, as it is the foundation of my self-worth. I am grateful to all further family members who nurtured my education, and my commitment to science and progress. My siblings have accompanied me my whole life, and are finishing their own doctorates as I write these sentences. We support each other, push each other, and are each other's safety nets when the world goes awry – I cherish our union. My partner is my most devoted supporter; I thank you for encouraging the effort and work that went into this thesis. Your loyalty is my guide.