SIIM
CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Eliminating Physician Dependency Bottlenecks in Automated Segmentation Quality Assurance Through a Novel Methodology

Reid Jockisch, OSF HealthCare; Matthew T. Bramlet, MD; Art Sedighi, PhD

Introduction/Background

Automated medical image segmentation transformed workflows by converting multi-hour manual annotations into fast, machine-generated models. These models are widely used in surgical planning and diagnostic applications. Despite improvements in segmentation accuracy, physician validation remains essential in validating clinical utility, creating a resource-intensive bottleneck that limits scalability and slows clinical integration. Current automated quality metrics, such as the Dice Similarity Coefficient (DSC), are insufficient during inference because they require comparison to ground truth, often unavailable in automated workflows.

Methods/Intervention

This study proposes a comparative quality assurance toolkit designed to reduce physician dependency by automatically flagging segmentation outputs that meet clinically relevant thresholds of accuracy. The approach leverages multiple independently trained machine learning models to generate predictions for the same imaging data. One model’s output is designated as a pseudo-ground truth, enabling computation of comparative metrics such as DSC, Jaccard Index, Average Symmetric Surface Distance, Hounsfield Distance, and Volumetric Similarity between model predictions. Metrics, paired with physician-assigned pass/fail labels, are used to train predictive models capable of classifying segmentation outputs as acceptable or requiring further review.

Results/Outcome

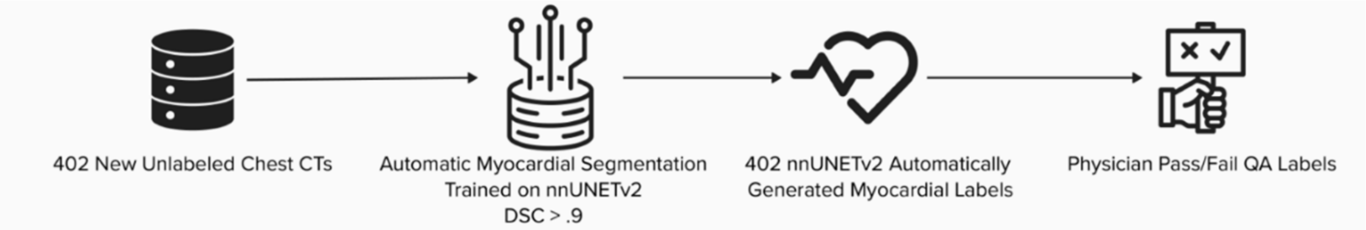
In a retrospective study, physician review of 402 segmentations revealed only 52.7% were clinically accurate, significantly lower than the assumed 90% suggested by DSC ($p< 0.001$), demonstrating that high DSC does not equate to clinical utility. A predictive QA model trained on comparative metrics achieved up to 82% accuracy and a low false-positive rate, approaching DSC’s performance without requiring ground truth. Comparative evaluation using DeLong’s and McNemar’s tests indicated no statistically significant difference in predictive performance between the proposed model and DSC on a holdout set ($p>0.05$), despite the predictive model capturing unique clinically relevant cases that DSC missed.

Conclusion

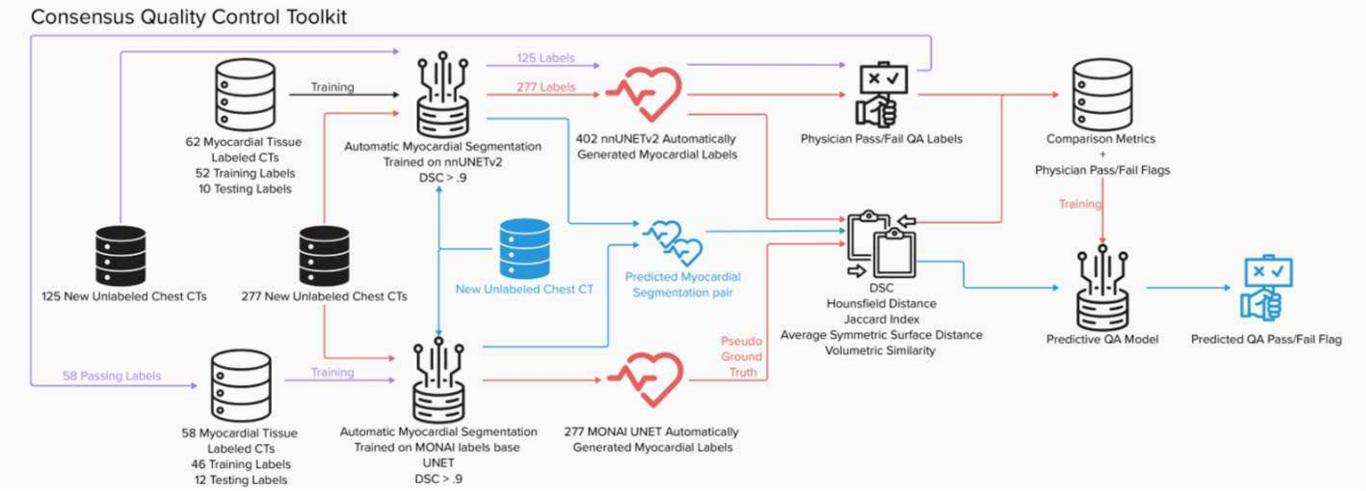
This work outlines a novel framework to partially automate QA in automated segmentation by aligning quality metrics with clinical relevance and reducing reliance on manual review.

Statement of Impact

The proposed methodology provides engineering leaders and clinical teams with a decision-making framework to implement scalable, clinically aligned QA systems. By automating the triage of low-risk segmentation outputs, it enables more efficient allocation of physician resources, accelerates surgical planning workflows, and supports broader adoption of automated segmentation technologies in clinical care.


402 New Unlabeled Chest CTs
Automatic Myocardial Segmentation
Trained on nnUNETv2
DSC > .9
402 nnUNETv2 Automatically
Generated Myocardial Labels
Physician Pass/Fail QA Labels

Simple experimental flow that demonstrates the process used to generate the clinical accuracy Pass and Fail flags used both in predictive model training, and comparison of the Dice similarity coefficient to Clinical utility.


Consensus Quality Control Toolkit
125 Labels
277 Labels
Training
62 Myocardial Tissue
Labeled CTs
52 Training Labels
10 Testing Labels
Automatic Myocardial Segmentation
Trained on nnUNETv2
DSC > .9
402 nnUNETv2 Automatically
Generated Myocardial Labels
Physician Pass/Fail QA Labels
Comparison Metrics
+
Physician Pass/Fail Flags
Training
125 New Unlabeled Chest CTs
277 New Unlabeled Chest CTs
New Unlabeled Chest CT
Predicted Myocardial
Segmentation pair
DSC
Hounsfield Distance
Jaccard Index
Average Symmetric Surface Distance
Volumetric Similarity
Predictive QA Model
Predicted QA Pass/Fail Flag
Pseudo
Ground
Truth
58 Passing Labels
Training
58 Myocardial Tissue
Labeled CTs
46 Training Labels
12 Testing Labels
Automatic Myocardial Segmentation
Trained on MONAI labels base
UNET
DSC > .9
277 MONAI UNET Automatically
Generated Myocardial Labels

A process map for the methodology explained by this research. This images shows how two independent datasets were used to train two automatic segmentation models, and how their results were correlated as training data for our novel QA predictive tool.

Keywords

Automated Segmentation; Quality Assurance; Machine Learning; Clinical Workflow; Dice Similarity Coefficient; Clinical Utility