Medical Image Analysis 116 (2027) 104351

Contents lists available at ScienceDirect

Medical Image Analysis

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

Ensemble-based uncertainty enables large-scale 3D segmentation quality control in population imaging

Tobias Norajitra [a,b,*], Tobias Nonnenmacher [d], Tobias Haueise [f,g], Fabian Isensee [c], Maximilian Zenk [a,x], Lea J. Michel [e], Jan Warszawski [a], Marc-Nicolas von Itter [e], Steffen Ringhof [e], Thoralf Niendorf [i], Jessica Krepel [j], Manuela Harries [j], Rafael Mikolajczyk [k], Jochen Hirsch [l], Daniel C. Hoinkiss [l], Thomas Kroencke [m,n], Josua Decker [m], Filippo Michelotti [o], Christoph M. Friedrich [p,y], Johannes Haubold [q], Tobias Pischon [r,s,t], Sigrid Thierry [u], Stefan N. Willich [v], Carsten O. Schmidt [w], Jakob Weiß [e], Susanne Rospleszcz [e], Hans-Ulrich Kauczor [d], Jürgen Machann [f,g,h], Fritz Schick [f,g,h], Christopher L. Schlett [e], Fabian Bamberg [e], Johanna Nattenmüller [e], Klaus Maier-Hein [a,b]

[a] Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany
[b] Pattern Analysis and Learning Group, Department of Radiation Oncology, University Hospital Heidelberg, Germany
[c] HIP Applied Computer Vision Lab, German Cancer Research Center (DKFZ), Heidelberg, Germany
[d] Department of Diagnostic and Interventional Radiology, University Hospital Heidelberg, Heidelberg, Germany
[e] Department of Diagnostic and Interventional Radiology, Medical Center - University of Freiburg, Faculty of Medicine - University of Freiburg, Freiburg, Germany
[f] Institute for Diabetes Research and Metabolic Diseases (IDM) of Helmholtz Munich at the University of Tübingen, Tübingen, Germany
[g] German Center for Diabetes Research (DZD), Neuherberg, Germany
[h] Section of Experimental Radiology, Department of Radiology, University Hospital Tübingen, Tübingen, Germany
[i] Berlin Ultrahigh Field Facility (B.U.F.F.), Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany
[j] Team Clinical Epidemiology, Department of Epidemiology (EPID), Helmholtz Centre for Infection Research, Braunschweig, Germany
[k] Institute for Medical Epidemiology, Biometrics, and Informatics, Interdisciplinary Center for Health Sciences, Medical Faculty of the Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
[l] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
[m] Department of Diagnostic and Interventional Radiology, University Hospital Augsburg, Augsburg, Germany
[n] Centre for Advanced Analytics and Predictive Sciences, Augsburg University, Augsburg, Germany
[o] Research Group Metabolic Imaging, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Germany
[p] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany
[q] Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany
[r] Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC), Molecular Epidemiology Research Group, Berlin, Germany
[s] Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC), Biobank Technology Platform, Berlin, Germany
[t] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[u] NAKO Study Center Augsburg, University Hospital Augsburg, Augsburg, Germany
[v] Institute of Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany
[w] Institute for Community Medicine, SHIP-KEF - Division Quality in the Health Sciences, University Medicine Greifswald, Germany
[x] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[y] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany

ARTICLE INFO

Keywords:
Segmentation quality control
Model uncertainty
Risk estimation
MRI
German National Cohort
UK Biobank

ABSTRACT

Modern Deep-Learning facilitates automatic segmentation in large imaging studies. To prevent segmentation errors from impacting data analysis, automated quality control (QC) for filtering segmentations and identifying errors (*segmentation selection*) is essential. QC must be fast and require minimal expert delineations for application to various ROIs in large datasets. Uncertainty quantification methods capable of estimating segmentation error without ground truth offer a promising foundation. However, their implicit error estimation may lack accuracy, making suitability for segmentation selection unclear. Unfortunately, commonly employed evaluation measures cannot capture their segmentation selection capability. Furthermore, selecting or rejecting segmentations based on uncertainty alone can be non-trivial. We propose a new method for efficient

* Corresponding author at: Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany.
E-mail addresses: t.norajitra@dkfz-heidelberg.de, t.norajitra@outlook.com (T. Norajitra).

https://doi.org/10.1016/j.media.2026.104351
Received 17 November 2025; Received in revised form 4 October 2026; Accepted 6 October 2026
Available online 14 October 2026
1361-8415/© 2026 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/).
Open access funding enabled and organized by Projekt DEAL.

segmentation QC in large imaging cohorts. We comprehensively evaluate uncertainty-based error estimators with respect to key goals of large-scale QC, specifically assessing segmentation selection capability through the area under the risk-coverage-curve. To optimize selection and rejection, we employ a binary search for decision thresholds, balancing the risk of including errors and unnecessarily rejecting segmentations. In > 11 000 MRI liver and lung segmentations in the German National Cohort, error estimation using the pairwise Dice of ensemble outputs achieved excellent QC, including segmentation selection. Numerous manually hard-to-detect errors and abnormalities ($n > 400$) were identified during high-quality result selection, measurably restoring integrity of volume and fat measurements. QC was fast ($0.29 - 4.48$ s/case) and required minimal annotated image data ($n < 125$). Our method can help tackle the challenge of reliable automated QC in large-scale segmentation studies.

1. Introduction

Large population-based cohorts like the German National Cohort (*NAKO*, Peters et al., 2022; Bamberg et al., 2015, 2024) or UK Biobank (*UKBB*, Littlejohns et al., 2020) offer a promising opportunity to investigate widespread diseases. They comprise data from magnetic resonance imaging (MRI), demographics, medical tests and questionnaires, that were collected from the general population in nation-wide multi-centric efforts.

Various studies have utilized the imaging data from these cohorts to investigate research questions related to different organs of interest. They focused, for instance, on body composition (fat, muscle volume) (Langner et al., 2019a; Küstner et al., 2020; Langner et al., 2021; Haueise et al., 2023; Wu et al., 2023), liver fat Whitcher et al. (2022), Jamialahmadi et al. (2023), cardio-vascular segmentation (Robinson et al., 2019; Bai et al., 2018; Wang et al., 2020; Ng et al., 2022; Mariscal-Harana et al., 2023), assessment of cardiac risk (Zheng et al., 2020; Ferdian et al., 2020; Elghazaly et al., 2023; Mahmood et al., 2023; Jamialahmadi et al., 2023) and cardio-vascular function (Mauger et al., 2019; Bai et al., 2020; Hepp et al., 2020; Pirruccello et al., 2022), kidney disease (Thanaj et al., 2023), pancreas fat, iron, T1-weighted signal intensity, and cancer (Basty et al., 2020; Triay Bagur et al., 2022; Chen et al., 2023), bone quality (Raisi-Estabragh et al., 2021; Fischer et al., 2021), ocular diseases (Alipanahi et al., 2021) or aging (Hepp et al., 2021).

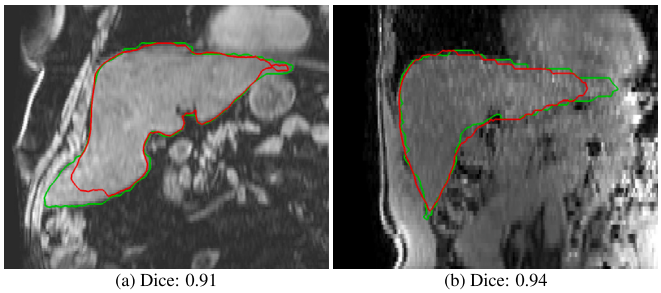
In these types of studies, image segmentation algorithms are typically employed to enable the targeted, spatial analysis of clinically relevant information within specific regions of interest. In this regard, the use of automatic segmentation methods based on deep learning has shown great success (Litjens et al., 2017; Isensee et al., 2021). Despite their effectiveness, however, these methods can sporadically introduce segmentation errors in the datasets to which they have been applied, potentially compromising downstream data analyses. To mitigate this, a dedicated segmentation quality-control (QC) step is typically performed, focusing on identifying and filtering such errors. In large population-based cohorts, manual error detection is highly time-consuming, making the implementation of automated segmentation QC techniques highly desirable.

So far, segmentation QC in NAKO or UKBB data has been approached in different ways. Some works have undertaken extensive visual quality checks, sometimes using specialized tools (Bai et al., 2018; Langner et al., 2019b, 2020; Bai et al., 2020; Liu et al., 2021; Thanaj et al., 2021; Kart et al., 2022; Wu et al., 2023; Thanaj et al., 2023). Other works have predicted segmentation accuracy directly from pairs of images and segmentation masks. In Robinson et al. (2017, 2019), a registration-based technique called reverse classification accuracy was successfully employed for that purpose. Without the need for prior training, high Pearson correlation coefficients ($r = 0.90$ in Robinson et al., 2017) between predicted and actual Dice values for cardiac segmentation in UKBB data were reported, with the only downside of increased computation times (11 min/case). In Robinson et al. (2018), segmentation accuracy was predicted on the same dataset using a convolutional neural network (CNN) trained with 12 880 segmentations, which were generated by applying a previously trained random forest (RF) to 2D slices from 4882 manually annotated 3D images. Faster computation was achieved (40 ms/case on a graphics processing unit, GPU) and, according to Wang et al. (2020), higher Pearson correlation ($r = 0.97$). In Wang et al. (2020), accuracy on the same dataset was predicted using a variational autoencoder (VAE) by reconstructing pairs of images and segmentations after performing a search in the latent space. Similar results were reported ($r = 0.96$) using less training data ($n = 1000$) and showing better generalization to another dataset. In Hann et al. (2021), a linear model was trained to predict the accuracy of aorta segmentations produced by a heterogeneous U-Net ensemble on UKBB data, using annotated 2D images ($n = 2121$) for training. Strong correlation was displayed efficiently on-the-fly following segmentation ($r = 0.90$).

Despite the strong ability of these methods to estimate segmentation accuracy, it remains unclear whether they can be easily applied for QC in various emerging scenarios. One challenge is that achieving the reported performance either required larger amounts of image annotations (Robinson et al., 2018; Wang et al., 2020; Hann et al., 2021) or longer runtimes (Robinson et al., 2017, 2019). Increased runtimes can become a significant burden in datasets containing tens of thousands of images. Additionally, for many use-cases and organs of interest, manual annotations are unavailable in such large quantities. A second challenge is that higher Dice values (> 0.90) were not perfectly predicted by the presented methods, a limitation also highlighted by Galati and Zuluaga (2021). However in a variety of scenarios, the Dice thresholds below which segmentations can be considered erroneous, lie precisely within this higher Dice range (see Fig. 1). Under these circumstances, low quality segmentations are likely to be erroneously approved by the presented methods, and it is not immediately clear how to cope with this issue.

Few works have investigated the use of uncertainty quantification for performing QC in UKBB, where neural network model uncertainty is leveraged for error estimation of predictions. In Langner et al. (2021), uncertainty of a CNN for regressing muscle and body fat from MR images was used for error estimation, based on the variance of the model output. Using uncertainty effectively allowed removing large errors from the produced results. Additionally, an analysis was presented on how filtering predictions based on their uncertainty would affect the overall prediction error and the number of remaining results. In Ng et al. (2022), various methods for measuring uncertainty of cardiac segmentations using 2D CNNs were examined for potential use in segmentation QC. A Spearman correlation coefficient of 0.69 between segmentation error and model uncertainty in UKBB was reported. Moreover, an analysis was presented how well the compared uncertainty measures were suited to suggest segmentations as potential candidates for manual correction. There, the area under the cases-reviewed-errors-remaining curve was considered for evaluation.

Using uncertainty-based error estimators for segmentation QC offers promise for a general use in large cohort data, since these estimators require no additional training and are highly time efficient. However in this scenario, their use entails a number of open problems. First, their error estimation reportedly demonstrated a weaker correlation with true segmentation accuracy on large cohort data (Ng et al., 2022), compared to explicit Dice prediction methods (Robinson et al., 2017, 2019, 2018; Wang et al., 2020; Hann et al., 2021). This raises the general



(a) Dice: 0.91 (b) Dice: 0.94

Fig. 1. Impact of segmentation errors on MRI-based measurements: Examples of falsified liver volume (a: 3087 ml segmented vs 3679 ml expected) and liver fat (b: 4.33% segmented vs 5.04% expected) in automatically segmented 3D regions (red) as compared to the ground truth (green), despite high Dice accuracy (> 0.90). These errors are a source for misdiagnosis of fatty liver disease and for false anatomical measurements from underlying populations.

question whether uncertainty-based approaches can successfully be employed in our context. Second, sensitivity and specificity are typically employed as performance measures to assess error detection capability as a sole aspect of segmentation QC. However, these measures fail to sufficiently reflect other important aspects of the broader task. For instance, when an error estimator is used in practice to divide a set of segmentations into erroneous (*set 1*) and error-free cases (*set 2*), these metrics reflect only the proportion of all true errors correctly included in set 1 (sensitivity), and the proportion of all error-free cases correctly included in set 2 (specificity). However, they do not explicitly inform about the relative risk of set 2 to contain errors, for which precision would have to be considered as an additional measure. Consequently, previous evaluations have not assessed how well error estimators support a crucial goal in large-scale segmentation QC: selecting accurate segmentations from an underlying set of results while minimizing both the risk of error and the rejection of good segmentations (*segmentation selection*). And third, for any suboptimal error estimator to achieve this particular goal, a decision must be made on how to accept or reject segmentations, establishing an optimal tradeoff between the risk of including errors and the total number of results obtained. However, no techniques or best practices have been proposed how to optimally make that decision in large cohort analysis.

In this work, we address the aforementioned limitations, proposing an efficient uncertainty-based approach for automated segmentation QC in large medical imaging datasets. We employ uncertainty quantification for error estimation, including techniques measuring the volumetric overlap of segmentation ensemble outputs (Roy et al., 2018; Hoebel et al., 2020), which have shown particular promise in smaller studies. We evaluate the proposed error estimators comprehensively, using specific performance measures to assess their suitability for three meaningful goals of large-scale segmentation QC: (1) to estimate the delineation accuracy of segmentations across the entire spectrum of the underlying dataset, (2) to specifically sample segmentation outliers for inspection and as potential input candidates for model refinement, and (3) to perform the key task of segmentation selection as described above. Finally, we propose an automated search for uncertainty thresholds to optimally select segmentations (Geifman and El-Yaniv, 2017), such that specific risk targets can be fulfilled while minimizing the unnecessary rejection of results.

Our main contributions are:

- A demonstration that ensemble-based uncertainty can serve as a proxy for segmentation error in large cohort datasets,
- Introduction of a comprehensive set of task-specific evaluation metrics (risk-coverage curves, precision–recall curves) that highlight selection and error-sampling capabilities,
- A threshold search procedure that allows tuning a desired risk level,
- A demonstration that the proposed QC approach isolates error cases at scale with minimal computational or labeling overhead, thereby restoring consistency in segmentation-derived measurements.

2. Related work

Various methods for automated error estimation and segmentation QC have been proposed in prior work. In the context of large-scale segmentation QC, these approaches are best discussed with respect to the following key aspects: the scalability of core error estimation techniques (Section 2.1), evaluation of dataset-level performance (Section 2.2), and the ability to tune failure detection for optimal real-world performance (Section 2.3).

2.1. Scalability of prior error estimation techniques

Computational quality control essentially relies on the ability to quantify existing errors in the absence of ground truth (*error estimation*). Various approaches have been proposed for this purpose. Some of them rely exclusively on model uncertainty as a proxy for error, without incorporating subsequent error-estimation steps (Roy et al., 2018; Hoebel et al., 2020; Langner, 2021; Ng et al., 2022). Roy et al. (2018) and Hoebel et al. (2020) relied on segmentation uncertainty to estimate error at the image-level. Estimates were obtained by aggregating foreground voxel-level uncertainties derived from Monte-Carlo dropout or ensembling. These uncertainties were computed based on either the maximum softmax probability or the entropy of softmax outputs. Alternatively, uncertainty was estimated from structural variation among segmentation outputs, measured via pairwise Dice discrepancy or the coefficient of variation of volumetric differences across output masks. Ng et al. (2022) employed additional approaches, including Bayesian neural networks, mutual information of softmax outputs, and pairwise surface distance between segmentation outputs, to derive both voxel- and image-level uncertainties. Langner (2021) derived image-level uncertainty of body composition estimates from the variance of Gaussian model outputs.

Other approaches estimate error by measuring the agreement between candidate segmentations and estimated ground-truth surrogates (Robinson et al., 2017, 2019; Wang et al., 2020). In Robinson et al. (2017, 2019), these surrogates are derived from atlas annotations after their registration to the input. Wang et al. (2020) use VAEs to generate hypotheses of ground-truth image-segmentation pairs.

A third group of approaches employs regression models to estimate delineation accuracy in the absence of ground truth. Some approaches estimate accuracy directly on image-segmentation pairs (Robinson et al., 2018; Billot et al., 2023; Qiu et al., 2023). Others leverage handcrafted features from images and segmentation masks (Kohlberger et al., 2012), segmentation uncertainty (Jungo et al., 2020; Hann et al., 2021; Kalkhof and Mukhopadhyay, 2023), or estimated ground-truth surrogates (Liu et al., 2019; Li et al., 2022) to inform the prediction. Beyond the uncertainty quantification and surrogate generation techniques described above, Jungo et al. (2020) proposed employing networks to predict data-intrinsic (aleatoric) uncertainty and to estimate error at the voxel-level. Furthermore, they investigated how informative the spatial distributions of various uncertainty and error estimates are for downstream failure detection and localization at both the image and dataset levels. Liu et al. (2019) and Li et al. (2022) proposed estimating segmentation and image surrogates, respectively, to inform error estimation. Different model types were employed to predict delineation accuracy, including support vector machines (Kohlberger et al., 2012), CNNs (Robinson et al., 2018; Li et al., 2022; Billot et al., 2023; Qiu et al., 2023), RFs (Jungo et al.,

2020), or linear models (Liu et al., 2019; Hann et al., 2021; Kalkhof and Mukhopadhyay, 2023).

To enable segmentation QC at scale, error estimation techniques need to be resource efficient, since long runtimes, or reliance on large training datasets hinder application across extensive data volumes and varying regions of interest (ROIs). At the same time, error estimation must align well with real underlying errors, to enable reliable error detection and minimize rejection of error-free results.

Prior work indicates that existing approaches face an inherent trade-off between resource efficiency and accurate alignment with real errors. Those approaches that demonstrated the strongest alignment with real error either raised significant training data requirements of >1000 annotated cases (Robinson et al., 2018; Wang et al., 2020; Hann et al., 2021; Billot et al., 2023; Qiu et al., 2023), or incurred increased computation times of ∼11 min per case (Robinson et al., 2017, 2019). Other approaches required less data during error estimation training, while still showing strong correlation with underlying error when validated on sizable test data (Kohlberger et al., 2012; Liu et al., 2019; Jungo et al., 2020). However, in direct comparison with the previously described methods, they exhibited weaker correlations, and the results in Liu et al. (2019) showed higher variance in the error estimates, with frequent overly optimistic predictions of accuracy. Finally, the purely uncertainty-based approaches - while being fast and not requiring estimator training - generally showed lower alignment with real errors compared to the other approaches (Roy et al., 2018; Hoebel et al., 2020; Ng et al., 2022). Nevertheless, results indicated that the pairwise Dice of the output segmentations could achieve good alignment with the true error in a number of datasets.

Altogether, when considering these findings, the question arises whether the more efficient error estimation approaches would perform sufficiently well for real-world applications in large-scale datasets. However, this question cannot be answered based on the commonly reported correlation coefficients alone. In practice, it is less relevant how well error estimators are in alignment with true error, but how well they can separate error-free from erroneous segmentations across the entire dataset. This underlines the importance of further evaluation strategies beyond measuring correlation with true error, to enable a more comprehensive assessment of practical performance.

2.2. Dataset-level performance evaluation

Beyond reporting correlation coefficients, several studies have examined the effectiveness of error estimators for failure detection (Kohlberger et al., 2012; Robinson et al., 2018, 2019; Hoebel et al., 2020; Jungo et al., 2020; Hann et al., 2021; Li et al., 2022; Kalkhof and Mukhopadhyay, 2023; Billot et al., 2023). Performance was reported using confusion matrices, accuracy, sensitivity, specificity, or the area under the receiver operating characteristic curve (AUROC), reflecting the achievable trade-off between detecting errors and correctly identifying error-free cases. These metrics capture the ability to make reliable decisions at the image level, that is, to classify individual segmentations as erroneous or error-free. For example, this can be relevant for on-the-fly assessment of cardiac MRI acquisition quality (Robinson et al., 2018).

In large-scale segmentation QC, however, image-level performance is not the primary concern (Section 1). Instead, the focus shifts to dataset-level decision making, where the goal is to select subsets of reliable, quality-controlled segmentations from the entire dataset. This requires assessing the purity of selected (presumably error-free) segmentations, captured by *precision*, in relation to their coverage across the dataset. Accordingly, evaluation must be performed at the dataset level rather than at the level of individual segmentations.

Only few works have considered QC performance in terms of such dataset-level objectives. Langner (2021) and Ng et al. (2022) evaluated error estimators with respect to reducing body-fat estimation errors and manual inspection and correction effort for segmentation failures, respectively, both assessed over the full dataset. Beyond these studies, no prior work has systematically investigated the use of error estimators in selecting subsets of reliable segmentations at the dataset level - a key requirement for enabling large-scale segmentation QC in practice.

2.3. Failure detection tuning

Tuning of failure detection is a crucial component of real-world segmentation QC, as it is required to achieve reliable performance in practice. It is typically performed by selecting an operating point (i.e., a failure detection threshold) for the error estimator, thereby balancing the trade-off between detection rates and false alarms.

Most approaches do not tune operating points for optimized failure detection. In methods that directly predict delineation accuracy (Section 2.1), the operating point of the failure detector is typically fixed at the predefined delineation accuracy threshold, below which segmentations are considered erroneous (Kohlberger et al., 2012; Robinson et al., 2017, 2018, 2019; Hann et al., 2021; Kalkhof and Mukhopadhyay, 2023). However, adjusting the operating point - for example, by shifting it to higher values - can reduce the number of undetected errors and thereby lead to more desirable outcomes.

Some works reported AUROC for image-level (Jungo et al., 2020; Billot et al., 2023) or pixel-level failure detection (Li et al., 2022), which can serve as a foundation for tuning. In this context, Jungo et al. (2020) and Billot et al. (2023) derived optimal operating-points from the receiver operating characteristic (ROC) curve, thereby tuning image-level failure detection. Additionally, Jungo et al. (2020) proposed an optimized thresholding of voxel-level uncertainties to prioritize information from erroneous regions of the segmentation.

However, segmentation selection performance at the dataset-level remains unexplored in prior studies (Section 2.2), and corresponding tuning strategies have not been investigated. Furthermore, approaches that rely exclusively on model uncertainty for error estimation lack a natural operating point, such as those based on predefined delineation accuracy thresholds. Their use for segmentation QC therefore requires the selection of a principled failure detection threshold. In this context, only Ng et al. (2022) and Langner (2021) presented tuning thresholds for uncertainty-based error estimators, although not for segmentation failure detection.

Another limitation of existing QC approaches is that performance tuning does not account for less obvious segmentation errors occurring at higher Dice values (Fig. 1(a)). Fixed Dice or Jaccard thresholds used to define errors span a wide range (0.6–0.9), sometimes even for the same ROI (Roy et al., 2018; Robinson et al., 2019; Hann et al., 2021). Only Kohlberger et al. (2012) considered stricter thresholds at higher values (≥ 0.9) of the more error-sensitive Jaccard index. As a result, subtle but relevant segmentation failures may go undetected by most existing QC approaches.

In summary, tuning failure detection is essential for practical segmentation quality control. However, only few prior works have addressed this topic, and only in the context of image-level performance. Performance tuning at the dataset-level - also for detecting less obvious errors - remains largely unexplored.

3. Novelty and contributions of the proposed method

As outlined in Section 2, current approaches face three key limitations that hinder their practical application to large-scale segmentation QC: substantial annotation or computational requirements for accurate error estimation (Section 2.1), evaluation strategies that do not assess dataset-level segmentation selection capability (Section 2.2), and a lack of corresponding tuning strategies (Section 2.3). These limitations are particularly relevant in the context of uncertainty-based error estimators. Such estimators are both fast and data-efficient, which makes them promising for truly scalable segmentation QC. However, current evaluation and tuning strategies do not establish whether they perform

sufficiently well for dataset-level segmentation selection or how to optimize them for this task.

With the proposed method, we address the aforementioned obstacles to large-scale segmentation QC. (1) We introduce a set of evaluation metrics specifically aimed at assessing performance at the dataset level, including the key task of segmentation selection. (2) We employ these metrics to identify efficient error estimation techniques that extend to large-scale applications and enable dataset-level QC. (3) We propose a performance tuning strategy to achieve desirable trade-offs at the dataset level for segmentation selection. And (4) we demonstrate, for the first time, the effectiveness of a data- and computation-efficient approach in retrieving high-quality segmentation results from real-world, large-scale population imaging data. To the best of our knowledge, no prior work has demonstrated comparable capabilities.

Methodologically, we introduce the following contributions to large-scale medical segmentation QC: an evaluation and optimization framework for uncertainty-based error estimators that lays the foundation for efficient large-scale segmentation QC; a dataset-level, selection-centric evaluation strategy and a mechanism for tuning segmentation selection performance, as key components; and a scalable, validated approach for retrieving optimal sets of quality-controlled segmentations from large datasets.

In this context, we provide practical solutions for effective error estimation using ensemble-based uncertainty, including robust aggregation of model outputs and interoperability with the high-performance segmentation baseline nnU-Net (Isensee et al., 2021). Furthermore, we introduce an initialization rule for the search of operating points during tuning to ensure reliable convergence. Finally, we provide empirical evidence for the effectiveness of the proposed error estimators and the overall QC approach using large-scale imaging data, encompassing diverse imaging conditions, separate ROIs, and visible pathologies.

Building on these contributions, we address the following key research question: Can scalable, resource-efficient error estimators be employed in practice, to retrieve faithful quality-controlled segmentations from large-scale data?

4. Methods

Our approach incorporates three main components: uncertainty-based error estimation in the absence of ground truth, task-specific evaluation of error-estimator performance, and automated threshold optimization for selecting reliable segmentations across the dataset. Together, these components provide a framework for identifying suitable error estimators and tuning their selection thresholds to balance residual error risk against dataset coverage.

4.1. Segmentation error estimation

Computational error-estimation techniques are typically employed in segmentation QC approaches, to estimate the presence and severity of errors in the underlying segmentations when ground truth is unavailable. Such quantitative error estimates are commonly defined within the $[0; 1]$ range, with higher values indicating greater estimated error. Once error estimates are obtained across a dataset of segmentations, they can be ranked by magnitude and paired with the corresponding segmentations. The quality of this ranking largely determines the achievable performance on specific QC tasks. For certain tasks, it is desirable that error estimates broadly reflect the actual degree of delineation inaccuracy in the segmentations. For others, it is more beneficial if few actual errors are present among segmentations assigned low estimated error values (cf. Section 4.2).

We propose leveraging segmentation model uncertainty to estimate error, aiming to enable fast, on-the-fly segmentation QC without the need for dedicated training. Specifically, we focus on ensemble-based uncertainty quantification methods (Roy et al., 2018; Hoebel et al., 2020; Section 1). Deriving error estimates from segmentation ensembles aligns well with the general goal of achieving high segmentation accuracy, as ensembling is commonly applied to enhance segmentation performance.

In this study, we compare a set of ensemble-based uncertainty quantification techniques for performing error estimation at the image-level. As a default, ensembles of 5 segmentation models are employed in all experiments, with the models trained on 5-fold cross-validation sets of the corresponding training data. Only binary ROI segmentations were considered, featuring one foreground and one background label. This includes lung segmentations where left and right lungs were treated as one single ROI. No postprocessing was performed on segmentations.

In two of the considered uncertainty quantification techniques (*max softmax, mean entropy*), error estimates are obtained by first computing voxel-level segmentation uncertainties separately for each ensemble model. These voxel-level uncertainties are then aggregated across the ensemble and over the segmentation foreground mask of the considered ROI. Specifically, the five uncertainty values per voxel are first averaged across the ensemble, yielding one average uncertainty value per voxel. These voxel-level averages are then averaged over the foreground mask, resulting in a single image-level error estimate. Foreground masks are obtained by averaging the softmax output maps of the five ensemble models, yielding one averaged probability map at the input-image resolution, followed by voxel-wise binarization using a threshold of 0.5 to obtain a single binary mask.

In two other techniques (*coefficient of variation, pairwise Dice*), error estimates are computed based on the binary segmentation masks derived from the individual ensemble models, by measuring discrepancies in mask overlap or mask-derived volume estimates.

The max softmax and mean entropy measures are calculated per model at each voxel based on respective softmax outputs. Specifically, in the max softmax measure, model uncertainty at voxel (u, v, w) is calculated from the highest softmax output $\boldsymbol{\sigma}(\mathbf{z})$ over the logits $\mathbf{z}$. For binary foreground–background segmentation, this can be represented implicitly using the foreground softmax output σ_1:

$$\mathrm{U}_{\mathrm{softmax}}(u, v, w) = 1 - \left|1 - 2\sigma_1(\mathbf{z}_{\mathrm{u,v,w}})\right| \tag{1}$$

This formulation assigns higher uncertainty to voxels whose foreground softmax probability is closer to 0.5. The deviation from 0.5 is inversely mapped to the $[0, 1]$ range, such that $\sigma_1 = 0.5$ corresponding to the maximum uncertainty value of $U = 1.0$. In the mean entropy measure, model uncertainty at voxel (u, v, w) is calculated as the entropy of the softmax output of the foreground and background classes $c \in \{0, 1\}$:

$$\mathrm{U}_{\mathrm{entropy}}(u, v, w) = -\sum_{c} \sigma_c(\mathbf{z}_{\mathrm{u,v,w}}) \log \sigma_c(\mathbf{z}_{\mathrm{u,v,w}}) \tag{2}$$

The other two measures (*coefficient of variation, pairwise Dice*) are calculated from the binary segmentation masks of each model in the ensemble. Given the final segmentation y generated by the model ensemble for a given image, the coefficient-of-variation uncertainty is computed from the mask voxel counts of the segmentations produced by the individual ensemble models. Specifically, the standard deviation σ of the voxel counts is normalized by their mean μ:

$$\mathrm{U}_{\mathrm{coefficient_of_variation}}(y) = \frac{\sigma_{\mathrm{mask_volume}}}{\mu_{\mathrm{mask_volume}} + \epsilon} \tag{3}$$

with ϵ added as a small constant for numerical stability. Similarly, the pairwise Dice (PWD) uncertainty for a final ensemble segmentation y is computed from the disagreement among the individual ensemble masks. Specifically, pairwise mask overlaps are quantified using the Dice coefficient and then converted into a disagreement measure. In the *mean PWD*, all pairwise Dice overlaps are first averaged into a single image-level value:

$$U_{\mathrm{mean_PWD}}(y) = 1 - \frac{1}{k} \sum_{i<j} \mathrm{Dice}(i, j) \tag{4}$$

where i, j index the segmentation masks produced by different ensemble models, and k denotes the number of unique mask combinations.

All possible pairwise combinations were included during calculation ($k = 10$ combinations for 5 ensemble model outputs), rather than randomly sampling mask pairs. The mean PWD from (4) may be sensitive to outliers from individual output masks, which do not necessarily indicate segmentation errors. Therefore, we compare it with the *median PWD*, which may reduce false positives in error estimation:

$$U_{\text{median_PWD}}(y) = 1 - \text{median}\left\{ \text{Dice}(i,j) \mid i < j \right\} \tag{5}$$

with the same variables i, j, and y as in (4). Since all error estimators are based on the same ensemble output masks, selecting a particular estimator for QC does not impact segmentation performance.

4.2. How to measure success in segmentation quality control

We evaluate the suitability of the proposed uncertainty-based error estimators for three key tasks in large-scale segmentation QC: (1) estimating delineation accuracy across the full spectrum of underlying segmentations, (2) selecting accurate segmentations for downstream analyses, and (3) sampling segmentation failures. To evaluate these tasks independently, we employ goal-specific performance measures.

Task 1 - Faithful assessment of delineation accuracy across the dataset

Reliable per-case prediction of delineation accuracy could directly support a range of segmentation QC tasks. However, obtaining such predictions reliably for each individual case is difficult in the absence of ground truth, particularly within the high-Dice range (Section 1). By contrast, in large-scale segmentation QC, it is also important to assess whether error estimates are globally aligned with the underlying distribution of segmentation accuracies across the dataset, rather than reliably calibrated for each individual case. Such alignment can characterize the overall distribution of segmentation quality, indicating the expected prevalence of high-quality segmentations as well as the frequency and distribution of errors and outliers in the dataset.

To this end, the Pearson correlation coefficient r is often used to evaluate the global alignment between error estimates and delineation accuracy on a set of available ground-truth data, such as that available during segmentation or error estimation training:

$$r = \frac{\text{cov}_{\text{accuracy,1-uncertainty}}}{\sigma_{\text{accuracy}}\sigma_{\text{1-uncertainty}}} \tag{6}$$

where cov denotes the covariance between accuracy and prediction confidence (= 1 − uncertainty), and σ represents their standard deviations. Thus, correlation coefficients closer to 1 indicate a stronger global alignment between error estimator outputs and delineation accuracy.

Task 2 - Maximized selection of accurate segmentations

As discussed in Sections 1 and 2.2, segmentation selection requires evaluation at the dataset level rather than through conventional image-level failure-detection metrics. Uncertainty-based error estimators perform segmentation selection by prioritizing the least uncertain cases in the dataset. Erroneous segmentations that are assigned low uncertainty may therefore be selected as reliable, resulting in *silent failures*. Such failures can compromise even restrictive selections that include only the most confident segmentations. Importantly, sensitivity and specificity are inadequate for assessing these failures.

As a solution to these problems, we translate ideas from the domain of image classification (Geifman and El-Yaniv, 2017) to the setting of large-scale segmentation QC. For our purposes, we assume a set of automatic segmentations S_m containing m pairs of images and predicted segmentations $(x, f(x)) = (x, y)$ using segmentation model f together with an uncertainty measure $U(x, y)$ for these predictions. Let $F(x, y) \in \{0, 1\}$ be an indicator function, evaluating to 1 if segmentation (x, y) is erroneous, and 0 otherwise. F can, for instance, be defined by a continuous segmentation error measure $e(y, y_{\text{gt}})$ between prediction y and ground truth y_{gt}:

$$F(x,y) = \begin{cases} 1, & e(y, y_{\text{gt}}) \geq \tau \\ 0, & \text{else} \end{cases} \tag{7}$$

Common segmentation metrics can be used to define $e(y, y_{\text{gt}})$, such as the Dice coefficient, the Jaccard similarity, or various surface distance metrics. A meaningful threshold τ can be heuristically determined by analyzing accuracy scores of typical segmentation errors on (cross-)validation sets. The threshold is generally chosen based on the clinical use-case for which the automatic segmentations are intended, considering the point at which errors become unacceptable for that specific application (see Section 4.4).

After choosing a threshold θ for uncertainty $U(x, y)$, we can select segmentations based on θ as defined by the selection function g_θ:

$$g_\theta(x,y) = \begin{cases} 1, & U(x,y) < \theta \\ 0, & else \end{cases} \tag{8}$$

Accordingly, we use $g_\theta(S_m)$ to denote the subset of segmentations as selected by g_θ. We can define the *selective risk* R as the expected risk of encountering errors in the selection:

$$R(f,g) = \frac{\mathbb{E}\left[F(x,y)g_\theta(x,y)\right]}{\phi(f,g)} \tag{9}$$

where coverage $\phi(f, g) = |g_\theta(S_m)|/m$ is the proportion of selected results out of all results. Typically, error risk is balanced against result coverage, meaning we can generally lower the risk of errors by selecting fewer but more certain segmentations.

In a scenario where the goal is to select quality-controlled segmentations for downstream analyses, the problem can be formulated as making a selection g_θ by choosing θ in Eq. (8), such that the selective risk R does not exceed a desired risk r^* of including errors, and at the same time, from all selections that fulfill this risk target, choosing the one that maximizes coverage ϕ (*optimization of the risk–coverage trade-off*). This explicitly formulates an objective that is often implicitly desired in the general setting of segmentation QC, such as in NAKO or UKBB - namely, predicting error cases and filtering results such that the remaining data have a low probability of error, while simultaneously minimizing the exclusion of valid results. So far, the problem has, however, not been stated or analyzed as such.

Unlike sensitivity and specificity, risk R and coverage ϕ are ideally suited to express an error estimator's ability to select segmentations, specifically by measuring the proportion of results that can be selected using threshold θ at a given risk of including errors. To allow a threshold-independent evaluation and comparison of segmentation selection performance, we analyze the risk-coverage (RC) curve on S_m. The RC curve is constructed by raising uncertainty threshold θ_i from low to high, while including the RC pairs of the resulting segmentation subsets $g_{\theta_i}(S_m)$. Consequently, the RC curve captures the overall segmentation selection performance of any error estimator by sampling the tradeoff profile for risk and coverage across all possible thresholds of θ. The area under the RC curve (*AURC*) can directly be employed as a performance measure to compare segmentation selection capability of different error estimators. Generally, the lower the AURC, the greater the coverage of results that can be achieved while keeping the risk of errors low.

One limitation of the AURC is its dependence on the error indicator function F, which defines failure cases. While such a definition is practically relevant in clinical settings, a fixed error criterion may also bias the evaluation of error estimators. For example, error estimators may appear inferior in segmentation selection if they include cases that fall just below the error threshold τ in Eqn. (7). To improve robustness in selecting error estimators for segmentation QC, we extend the proposed evaluation framework to include a continuous, threshold-agnostic variant of the risk-coverage curve alongside the threshold-dependent formulation. Specifically, we replace the binary error indicator $F(x, y)$ in Eq. (9) with the continuous error measure $e(y, y_{\text{gt}})$:

$$R_{\text{cont}}(f,g) = \frac{\mathbb{E}\left[e(y,y_{\text{gt}})g_\theta(x,y)\right]}{\phi(f,g)} \tag{10}$$

We incorporate both variants of the RC curve to evaluate the segmentation selection performance of error estimators.

Task 3 - Sampling of failure cases

Sampling typical failure cases at the high-uncertainty end of error estimator outputs is highly relevant in practice. It provides valuable insights into the extent and causes of segmentation errors within the dataset. Unlike selecting accurate segmentations, failure sampling focuses on identifying errors. Specifically, it aims to maximize the number of detected errors out of all errors in the dataset while maintaining high precision. Recall (sensitivity) directly reflects the proportion of failures in the dataset that are detected.

Accordingly, analyzing the precision–recall (PR) curve serves as an adequate technique for evaluating the effectiveness of error estimators in sampling failure cases (Ahmed and Courville, 2020). For the PR curve, we sample cases starting from the highest to the lowest error estimates. The goal of any error sampling is to reach a high recall of error cases ($Recall_F$) in the dataset:

$$Recall_F = \frac{TP}{P} \tag{11}$$

where TP represents the sampled true failure cases, and P denotes the total number of failure cases in the dataset. Simultaneously, it is ideal to maintain high precision in the sampling of errors:

$$Precision_F = \frac{TP}{TP + FP} \tag{12}$$

where FP represents the sampled non-failure cases. We construct the PR curve by progressively adding cases to the sample set and updating precision and recall, thus forming the curve. The PR curve provides insight into how effectively error cases can be sampled from the high end of error estimates, while minimizing the inclusion of error-free segmentations. Repeated inclusion of false positive cases at high error estimation values is heavily penalized by the precision component.

The area under the precision–recall curve ($AUPRC_F$) can be used to evaluate the effectiveness of error estimators in failure sampling, with higher $AUPRC_F$ indicating improved sampling performance. While related, AURC and $AUPRC_F$ capture complementary aspects of performance (cf. Appendix B.1).

4.3. Optimal threshold search for segmentation selection

In the previous section, we described the use of AURC as a measure for segmentation selection performance of error estimators. To actually use a specific error estimator in practice for performing segmentation selection, a selection threshold θ_{opt} (Eq. (8)) must first be determined based on a limited set of ground-truth annotations. This threshold should meet the desired risk target r^* through its selection while maximizing the result coverage, thereby optimizing the RC trade-off (Section 4.2).

Choosing optimal thresholds in this context is non-trivial: for instance, when uncertainty thresholds are used to filter out segmentations above a certain threshold, increasing these thresholds can expand the set by including both erroneous and error-free cases. Thus, the error risk does not behave monotonically when thresholds change, which necessitates a dedicated search for an optimal solution.

Another challenge concerns the empirical estimation of error risk in the subsets $g_\theta(S_m)$ across the RC curve. These subsets represent random samples of segmentations and of errors as produced by segmentation method f. Consequently, the error risks of the RC curve are likewise empirical estimates of the true error rates inherent to f. Thus, for any threshold θ_i with corresponding error risk $R_i < r^*$ on the RC curve, when applied for segmentation selection in another dataset outside S_m (e.g., a target dataset without ground-truth annotations), it is unclear whether the actual error rate in the resulting selection will satisfy r^*, since the true error may have been underestimated in S_m.

To address the latter problem, based on the empirical errors in the RC curve, we can determine an upper bound on the error risk, following the proposition in Geifman and El-Yaniv (2017). That is, for any selection of segmentations, we determine the error rate of a worst case segmentation method - one whose performance is sufficiently inadequate that it would likely produce more errors than those observed. For any selected set $g_\theta(S_m)$ of segmentations produced by a selection function g_θ using uncertainty threshold θ (Section 4.2), we can determine upper risk bound p based on the observed empirical error risk $\hat{r}$ on the k segmentations of subset $g_\theta(S_m)$:

$$\sum_{i=0}^{\hat{r}k} \binom{k}{i} p^i (1-p)^{k-i} = \delta \tag{13}$$

Here, δ is a selectable residual risk for bound p not being valid, or conversely, $1-\delta$ being a guarantee for the validity of p. The bound p can be proven to be the tightest possible risk bound for errors in subset $g_\theta(S_m)$ (Geifman and El-Yaniv, 2017). Consequently, the goal of our uncertainty threshold search is to identify an optimal threshold θ^* such that, for the selected subset $g^*(S_m)$, the upper risk bound p^* is below an acceptable risk target r^*, while maximizing the coverage of results across the entire dataset. This can be expressed in terms of selective risk from Eq. (9):

$$P_{S_m}\{R(f, g^*) > r^*\} < \delta \tag{14}$$

To find optimal thresholds, we employ a binary search as proposed in Geifman and El-Yaniv (2017). Here, for the actual residual risk of δ to still apply over the course of the entire search, it must be aggregated for a total number of $\log N$ search iterations given sample size N. Thus, the residual risk for which p is determined at each iteration, is additionally corrected by a factor of $1/\log N$.

In practice, the proposed search allows us to explore segmentation selections available at varying risk levels. These selections can then be assessed based on their RC tradeoffs, which determine their suitability for specific downstream analyses. Determining which tradeoffs are suitable for specific analyses strongly depends on the nature of those analyses and must be evaluated on a case by case basis. This is something the proposed threshold search is designed to facilitate.

4.4. Deployment on large-scale data

Following the evaluation of candidate error estimators and threshold optimization for the selected candidate, a performance-tuned estimator is obtained that can be applied to large-scale imaging data in the absence of ground truth. During deployment, quantitative error estimates are generated from segmentation masks across the entire dataset. The previously optimized threshold is then applied to these estimates to assign a selection or rejection decision to each case. This enables the identification of a reliable set of quality-controlled segmentations alongside filtered failure cases (Fig. 2), thereby replacing the manual review of thousands of automatically generated segmentations.

Two parameters are specified for suitable application scenarios: τ for defining failure cases (Eq. (7)), and r^* for defining the target risk (Eq. (14)). τ is typically chosen based on the underlying ROI and the segmentation accuracy required for the intended application (Section 1; comp. scenarios below) and is independent of the employed error estimator. r^* is inherent to the proposed method and provides a QC target without requiring manual selection of a less interpretable uncertainty threshold θ (Eq. (8)). The choice of r^* depends on the risk-coverage tradeoffs achievable with a given segmentation performance in the specific application scenario. Stricter QC targets may be selected when segmentation performance still permits sufficient coverage of the underlying dataset. Conversely, less strict targets may be appropriate when the benefit of increased coverage outweighs the corresponding increase in worst-case risk. Two general example application scenarios may be considered:

Scenario 1 – Dataset-specific deployment. The objective is to perform QC for the curation and analysis of an existing large-scale dataset. τ, and r^* are selected to satisfy requirements of downstream applications. This selection can be performed either heuristically or analytically, for

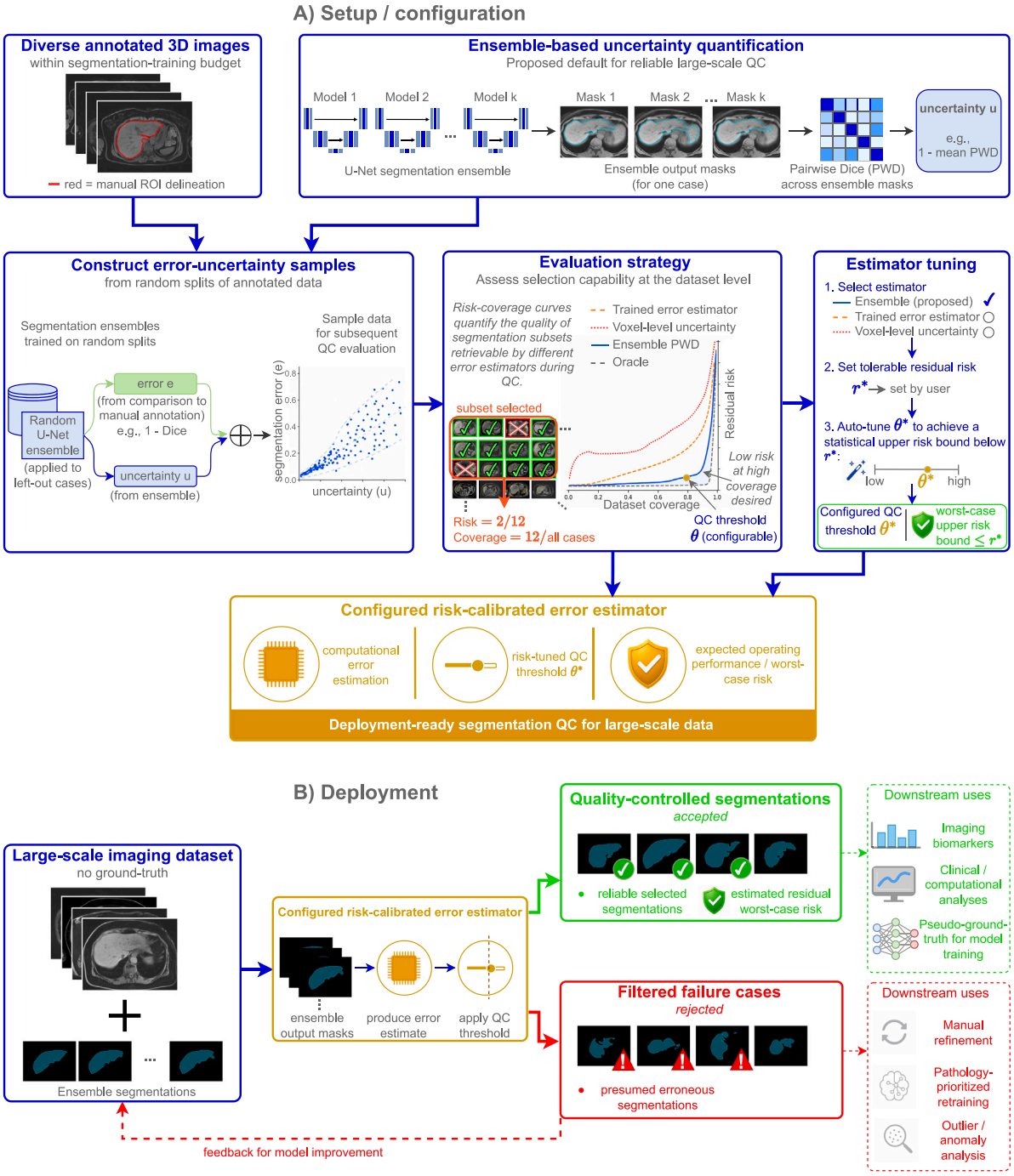


Fig. 2. Overview of the proposed method for large-scale segmentation quality control (QC): An evaluation and tuning framework (A) is introduced to compare and tune candidate error estimators for large-scale segmentation QC. Starting from limited ground truth data (A, top left), segmentation models are trained on randomly selected splits to induce a range of segmentation errors, which are paired with corresponding outputs from the candidate error estimators to generate evaluation samples (A, middle left). These evaluation samples are used to assess how effectively different estimators identify reliable subsets of segmentations across the full dataset (dataset-level evaluation; A, middle center). Among the evaluated approaches, uncertainty quantification using pairwise Dice discrepancies between ensemble segmentation outputs is proposed as a strong default strategy for segmentation QC (A, top right). In a final tuning step (A, middle right), the operating point of the best-performing estimator is automatically optimized under a statistical error risk bound. This risk-constrained tuning maximizes the number of accepted segmentations during QC while keeping the worst-case risk of residual errors among accepted segmentations below a user-defined tolerance. The risk-calibrated estimator (A, bottom) is deployed on large-scale imaging data in the absence of ground truth (B, left). During deployment, the estimator evaluates segmentation masks across the entire dataset and returns an optimized set of reliable, quality-controlled segmentations (B, top right) alongside filtered failure cases (B, bottom right). The filtered cases may then be used for large-scale outlier analysis or model retraining after manual refinement, including challenging pathological cases.

example by using sample data to quantify the impact of these parameters on clinical stratification performance, the consistency of imaging biomarker distributions, or the performance of downstream models trained on the resulting quality-controlled data. During QC, rejected failure cases that are clinically relevant are manually reviewed and reinserted into the quality-controlled dataset after manual refinement and, optionally, segmentation model retraining. This step helps prevent the under-representation of relevant cases in the final dataset.

Scenario 2 – Continuous deployment for general purpose QC. The objective is to provide QC for a continuous stream of imaging segmentations, for example those generated prospectively for a given imaging modality. Conservative values for τ, and r^* are selected based on the performance spectrum of the error estimator to ensure data quality suitable for a broad range of downstream applications. Rejected failure cases are manually reviewed and, if appropriate, refined for use in retraining the primary segmentation model or in training expert models that support the automatic segmentation of pathologies and edge cases.

5. Experimental setup

We evaluated our approach in liver and lung segmentation settings using data from the public NAKO (Bamberg et al., 2024), COSYCONET (Karch et al., 2016), and HELENA (Schübel et al., 2016) cohorts. Our experimental setup included several components: segmentation model training, error-estimator evaluation using segmentation errors obtained from randomly trained model ensembles, comparison with learning-based methods, and large-scale application to the full NAKO and COSYCONET datasets. The following subsections describe the corresponding data, models, and evaluation procedures.

5.1. Imaging data

5.1.1. German National Cohort (NAKO)

In this paper, we utilized large-scale MRI data from NAKO. Images were acquired from participants at five sites in Germany (n = 11 190) until 31st Dec 2016, using 3T MRI (MAGNETOM Skyra, Siemens Healthineers AG) (Bamberg et al., 2015, 2024). For liver segmentation and subsequent fat content quantification, we utilized axial images from T1-weighted 3D VIBE two-point (3.0 mm slice thickness, 1.40 mm× 1.40 mm in-plane spatial resolution) and six-point (4 mm slice thickness, 1.64 mm×1.64 mm in-plane spatial resolution) Dixon MRI. T1-weighted Dixon MRI is an imaging technique capable of separating fat and water signals, thereby enabling their quantification. The specific protocol used in the NAKO for two-point (2pt) Dixon MRI provides enhanced spatial resolution and tissue contrast compared to six-point (6pt) Dixon MRI, although it is less accurate for fat quantification. In clinical practice, 2pt-Dixon images are more commonly available. In our study, we utilized the images from both Dixon techniques to evaluate segmentation QC across two distinct qualities of imaging. Additionally, we aimed to analyze how QC affects the validity of imaging-derived volume and fat measurements. This impact can be effectively assessed through a joint analysis of 2pt- and 6pt-Dixon data.

5.1.2. HELENA

For segmentation model pretraining (see Section 5.3), we utilized MR data from the HELENA study (Schübel et al., 2016), which investigated dietary weight loss interventions among 150 overweight or obese participants without diabetes. 2pt- and 6pt-Dixon MR images were acquired at Heidelberg University Clinic using a 1.5T MR scanner (MAGNETOM Aera, Siemens Healthineers AG), with image resolution matching that of NAKO.

5.1.3. German COPD and Systemic Consequences - Comorbidities Network (COSYCONET)

Beyond the NAKO and HELENA datasets, we incorporated MRI data from the COSYCONET cohort (Karch et al., 2016). Images were acquired between 2010 and 2013 from participants across 14 sites in Germany (n = 598), using 1.5T MRI (MAGNETOM Aera, Siemens Healthineers AG). For lung segmentation and automated QC, we utilized axial images from T1-weighted 3D VIBE MRI (4.0 mm slice thickness, 1.25 mm × 1.25 mm in-plane spatial resolution), acquired using a fat-saturation protocol comparable to the Dixon technique (Konietzke et al., 2025). Additionally, we included paired computed tomography (CT) acquisitions with available ground-truth lung masks for the same subjects, to assess the impact of QC on volume measurements and to highlight cases in which subtle errors may remain undetected by volumetric measures. Inspiratory MRI and CT images were used in these experiments. After import from Digital Imaging and Communications in Medicine (DICOM) files and exclusion of cases with geometric mismatches, a total of 528 cases remained for analysis. The COSYCONET cohort comprises fat-saturated MR images acquired on different scanner hardware and at acquisition sites distinct from those contributing to the NAKO and HELENA cohorts. Furthermore, in contrast to NAKO and HELENA, the cohort exhibits higher prevalence of pathological cases, particularly COPD.

5.2. Ground-truth annotations

For pretraining the liver segmentation models, manual delineations of the liver had been performed by a medical expert on 2pt-Dixon images (n = 58) from the HELENA study. Annotations were made using interactive labeling tools from the Medical Imaging Interaction Toolkit (MITK, Wolf et al. (2004)), following a specifically developed standard operating procedure (SOP) for liver segmentation. After pretraining, additional training cases from NAKO were manually annotated by correcting faulty segmentations over multiple rounds as described in Section 5.3. In this process, n = 35 and n = 65 manual liver segmentations were produced by four medical experts for 2pt- and 6pt-Dixon images in NAKO, following the aforementioned standard operating procedure for liver segmentation, using interactive labeling tools from the NORA (Anastasopoulos et al., 2017) and MITK frameworks. An additional test set of n = 50 randomly selected 2pt-Dixon cases from NAKO was annotated by one medical expert on the NORA platform. In this case, manual liver annotations did not adhere to the previously established standard operating procedures. All dataset sizes were selected empirically, balancing observed model performance and test case representativeness with the cost of manual annotations.

For the lung segmentation experiments, manual lung annotations were performed by two medical experts. An initial set of n = 16 annotations was produced on NAKO data, followed by three rounds of manual correction addressing n = 7 faulty segmentations each, resulting in a total of n = 37 annotations. Additionally, n = 13 manual annotations were created for random cases from the held-out COSYCONET test set.

5.3. Deep segmentation

For the image segmentation task, we used nnU-Net (Isensee et al., 2021), a high-performance method known for its strong generalizability across various organs and imaging modalities in its 3D full-resolution configuration. To optimize liver segmentation performance on NAKO data, we first pretrained nnU-Net using the 58 available 2pt-Dixon image annotations from the HELENA study. This was followed by a repeated refinement on NAKO data, during which candidate segmentations were manually reviewed, corrected and added as training cases. Separate models were trained for the 2pt-Dixon and 6pt-Dixon images in NAKO. After pretraining on 2pt-Dixon images from the HELENA study (n = 58), the 2pt-Dixon model was refined on $n = 5\times7$ cases from

Table 1
Sampling of segmentation-uncertainty pairs for evaluation of the proposed error estimators: Segmentation ensembles were randomly trained on annotated 2pt- and 6pt-Dixon data using decreasing training set sizes, and applied on cases that were left-out during training each time. Uncertainties were calculated on the resulting segmentation masks using the proposed error estimators. Samples were additionally generated on separate test datasets. Counts vary due to random split sizes; identical values occur coincidentally.

Dataset / experiment	Train split size [%]	Trained ensembles	Obtained sample pairs
$NAKO_EXP_{2pd_liver}$	80.00	17	315
	50.00	8	371
	20.00	5	372
	10.00	9	N/A*
$NAKO_EXP_{2pd_test}$	–	17	850
	–	8	400
	–	5	250
	–	9	900
$NAKO_EXP_{6pd_liver}$	80.00	15	195
	50.00	6	195
	33.00	4	195
$NAKO_EXP_{2pd_lung}$	80.00	40	249
	50.00	16	256
	20.00	10	256
$COSYCONET_EXP_{test}$	–	40	520
	–	16	208
	–	10	130

* These ensembles were only applied to the separate $NAKO_EXP_{2pd_test}$ dataset; the resulting sample counts are reported there.

NAKO, and the 6pt-Dixon model was refined on $n = 23+6\times7$ cases from NAKO. Manual corrections were applied as described in Section 5.2. Model training was performed from scratch with each refinement. A default nnU-Net ensemble of five models was used, trained on cross-validation folds of the training data. Similarly, a lung segmentation ensemble was trained on $n = 16 + 3 \times 7$ NAKO cases.

5.4. Acquiring error estimation samples

For the comparative evaluation of the proposed error estimators, we generated a diverse set of segmentation model outputs, reflecting varying degrees of accuracy and error on our annotated images (Section 5.2). To achieve this, we applied a repeated random cross-validation procedure, training nnU-Net on randomly selected training subsets of progressively smaller sizes, as shown in Table 1. In each iteration, the trained models were applied to perform segmentation on the images that were not part of their respective training sets during that round. On these segmentation outputs, we calculated the uncertainties for the compared error estimators, generating a diverse set of segmentation-uncertainty pairs for subsequent analysis. Separate sets were generated for 2pt- and 6pt-Dixon data ($NAKO_EXP_{2pd_liver}$, $NAKO_EXP_{2pd_lung}$, $NAKO_EXP_{6pd_liver}$), comprising $n = 1058$, $n = 761$, and $n = 585$ samples, respectively (see Table 1). During liver model training for both Dixon sequences, annotated 2pt-Dixon cases from the HELENA study were integrated, in alignment with the training procedure outlined in Section 5.3. However, no segmentation-uncertainty samples from 2pt-Dixon cases in HELENA were part of the $NAKO_EXP_{6pd_liver}$ dataset.

The trained 2pt-Dixon models were used to generate additional samples ($NAKO_EXP_{2pd_test}$, $COSYCONET_EXP_{test}$, $n = 2400$, and $n = 858$, respectively) on the separate 2pt-Dixon and COSYCONET test sets described in Section 5.2. Obtaining erroneous samples was more challenging on $NAKO_EXP_{2pd_test}$. To address this, we intentionally decreased segmentation performance during inference, by disabling test time augmentation in nnU-Net and reducing the sliding window patch overlap to zero. Furthermore, models were trained on a further reduced training set size (see Table 1, line 4). These models were then applied both with and without the intentional performance reduction.

We defined an indicator function, F, to distinguish between error-free and erroneous segmentations in these datasets (Section 4.2). F was defined based on the delineation accuracy of the segmentations, measured using the Jaccard similarity coefficient $J(S_{\text{pred}}, S_{\text{ref}}) = \frac{|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{pred}} \cup S_{\text{ref}}|}$ between the predicted segmentation S_{pred} and the ground truth S_{ref}. We heuristically determined Jaccard thresholds below which segmentations were considered erroneous (detailed at the end of the paragraph). Liver segmentations in 6pt-Dixon exhibited a higher degree of labeling uncertainty due to low imaging contrast, necessitating different thresholds for the two MR sequences. Moreover, manual annotations in the additional test set were created using a different set of standard operating procedures compared to the primary dataset, resulting in systematic deviations from those in the primary dataset (Section 5.2). Specifically, adverse decisions regarding the inclusion and exclusion of the vena cava and portal vein cavity introduced systematic bias (see Appendix C.3 for corresponding inter-rater comparisons). To establish appropriate accuracy thresholds under these conditions, we visually inspected error cases across all datasets and compared their respective Jaccard similarities. Based on these comparisons, Jaccard thresholds of 0.92, 0.91 and 0.87 were determined for the liver in the 2pt-Dixon experimental, the 2pt-Dixon test and the 6pt-Dixon datasets, respectively. A Jaccard threshold of 0.95 was determined for the lung segmentations in the NAKO and COSYCONET datasets.

5.5. Comparison baselines

We compared the proposed uncertainty-based error estimators with more complex, learning-based methods, including a CNN-based regression approach following the works of Robinson et al. (2018) and Billot et al. (2023); a VAE-based approach following the work of Liu et al. (2019); and, a random-forest (RF) approach following Jungo et al. (2020). We trained all error estimators under the same low-data regime available in this project.

Specifically, we performed five-fold cross-validation with subject-level splits on the $NAKO_EXP_{2pd_liver}$, $NAKO_EXP_{6pd_liver}$, and $NAKO_EXP_{2pd_lung}$ datasets. We trained each method on the error samples from the training folds and evaluated on the corresponding test folds. For inference on the $NAKO_EXP_{2pd_test}$ and $COSYCONET_EXP_{test}$ datasets, we ensembled the fold models by averaging their predictions. This approach outperformed single models trained on the full training set.

The CNN approach predicted Dice accuracy from image-segmentation pairs, the VAE estimated segmentation quality based on mask reconstruction, and the RF predicted Dice accuracy from segmentation- and entropy-derived features. For the RF, radiomics- and heuristic-based feature variants were compared, with the better-performing heuristic variant used in subsequent experiments. Detailed preprocessing, architectures, training procedures, and hyperparameter optimization are provided in Appendix A.1.

5.6. Segmentation quality-control experiments

We conducted five experiments to evaluate the proposed segmentation QC (Expt. 1-5). First, we assessed the effectiveness of the compared uncertainty-based error estimators in performing the designated segmentation QC tasks. Specifically, we evaluated their performance in assessing delineation accuracy (Expt. 1), selecting accurate segmentations (Expt. 2), and sampling failure cases (Expt. 3). These experiments were conducted on the $NAKO_EXP_{2pd_liver}$, $NAKO_EXP_{6pd_liver}$, $NAKO_EXP_{2pd_test}$, $NAKO_EXP_{2pd_lung}$, and $COSYCONET_EXP_{test}$ datasets.

For the two top-performing error estimators from Expt. 2 (on 2pt- and 6pt-Dixon data), we evaluated the proposed optimal threshold search (Expt. 4). Optimal thresholds for selecting segmentations were

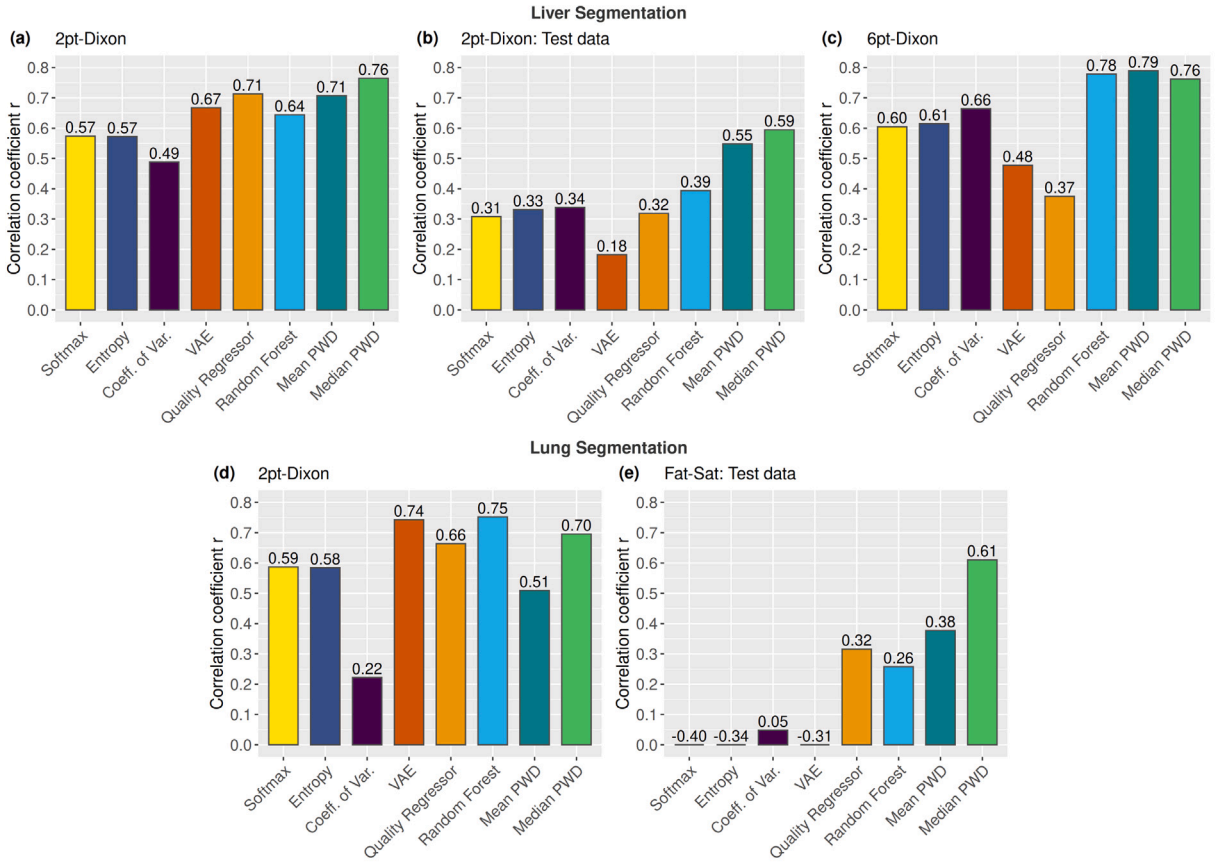


Fig. 3. Alignment of error estimator outputs with true segmentation error: PWD-based error estimators (green, cyan) showed the strongest overall alignment with liver (top row) and lung (bottom row) segmentation errors across 2pt-Dixon (a, b, d), 6pt-Dixon (c), and Fat-Sat data (e), as measured by the Pearson correlation coefficient. An exception was observed for the 2pt-Dixon lung segmentation experiment, where the RF-based approach achieved the highest correlation. In all other cases, PWD-based estimators consistently demonstrated stronger alignment with the true error than the softmax-, entropy-, coefficient of variation-, CNN-, VAE-, and RF regression-based error estimators (yellow, blue, purple, orange, red, and light blue).

determined on the $\text{NAKO_EXP}_{\text{2pd_liver}}$, $\text{NAKO_EXP}_{\text{6pd_liver}}$, $\text{NAKO_EXP}_{\text{2pd_lung}}$, and $\text{COSYCONET_EXP}_{\text{test}}$ datasets, ensuring that error risk remained within predefined target risk levels. For 2pt-Dixon, the computed risk bounds were further validated on the $\text{NAKO_EXP}_{\text{2pd_test}}$ and $\text{COSYCONET_EXP}_{\text{test}}$ datasets.

Finally, the selected error estimators and calculated thresholds were applied for segmentation selection across the whole NAKO and COSYCONET datasets (Expt. 5). Segmentation selection was performed on 2pt- and 6pt-Dixon segmentations generated by the three nnU-Net ensembles (detailed in Section 5.3). Liver volume, liver fat, and lung volume measurements were derived from both selected and rejected segmentations. In 11 172 NAKO cases and 528 COSYCONET cases, paired volume and fat measurements enabled joint analysis, as 2pt- and 6pt-Dixon as well as Fat-Sat MRI and CT images were available, respectively. In this analysis, we assessed the impact of segmentation errors on the consistency of paired measurements and the improvements achieved through error rejection with the proposed segmentation QC.

6. Results

Sections 6.1–6.6 present the results of Experiments 1-5, as outlined in Section 5.6.

6.1. Estimation of delineation accuracy

To assess the alignment between the error estimates produced by the compared estimators and delineation accuracy on our error estimation datasets, we quantified this relationship using the Pearson correlation coefficient. Delineation accuracy was measured by the Jaccard similarity (Section 4.2).

The resulting Pearson correlation coefficients are shown in Fig. 3. Overall, the PWD-based error estimators showed the most consistent strong agreement with delineation accuracy across ROIs and datasets, demonstrating the strongest correlations in most settings, specifically the median PWD on $\text{NAKO_EXP}_{\text{2pd_liver}}$, $\text{NAKO_EXP}_{\text{2pd_test}}$, and $\text{COSYCONET_EXP}_{\text{test}}$ ($r = 0.76$, $r = 0.59$, and $r = 0.61$, respectively), compared with the strongest non-PWD estimators in the same settings ($r = 0.71$, $r = 0.39$, and $r = 0.32$, respectively), as well as the mean PWD on $\text{NAKO_EXP}_{\text{6pd_liver}}$, where it showed the strongest correlation together with the RF ($r = 0.79$ and $r = 0.78$, respectively). In $\text{NAKO_EXP}_{\text{2pd_lung}}$, the median PWD achieved a near-strongest correlation, ranking slightly behind the RF and VAE ($r = 0.70$, $r = 0.75$, and $r = 0.74$, respectively). Other uncertainty-based estimators also reached substantial correlations in selected settings, but generally remained below the PWD-based measures. Participant-level bootstrap comparisons further supported these findings, with detailed paired performance differences and 95% confidence intervals reported in Appendix Fig. C.1 and Tab. C.1.The learned estimators showed greater sensitivity to the underlying ROI and dataset. Although some performed strongly in selected settings, including RF on $\text{NAKO_EXP}_{\text{6pd_liver}}$ as well as RF and VAE on $\text{NAKO_EXP}_{\text{2pd_lung}}$, their correlations were generally less stable across settings. For example, the VAE and quality regressor ranked among the lowest-performing methods on the $\text{NAKO_EXP}_{\text{2pd_test}}$ ($r = 0.18$ and $r = 0.32$, respectively) and $\text{NAKO_EXP}_{\text{6pd_liver}}$ ($r = 0.48$ and $r = 0.37$, respectively) datasets, and the VAE correlated negatively on

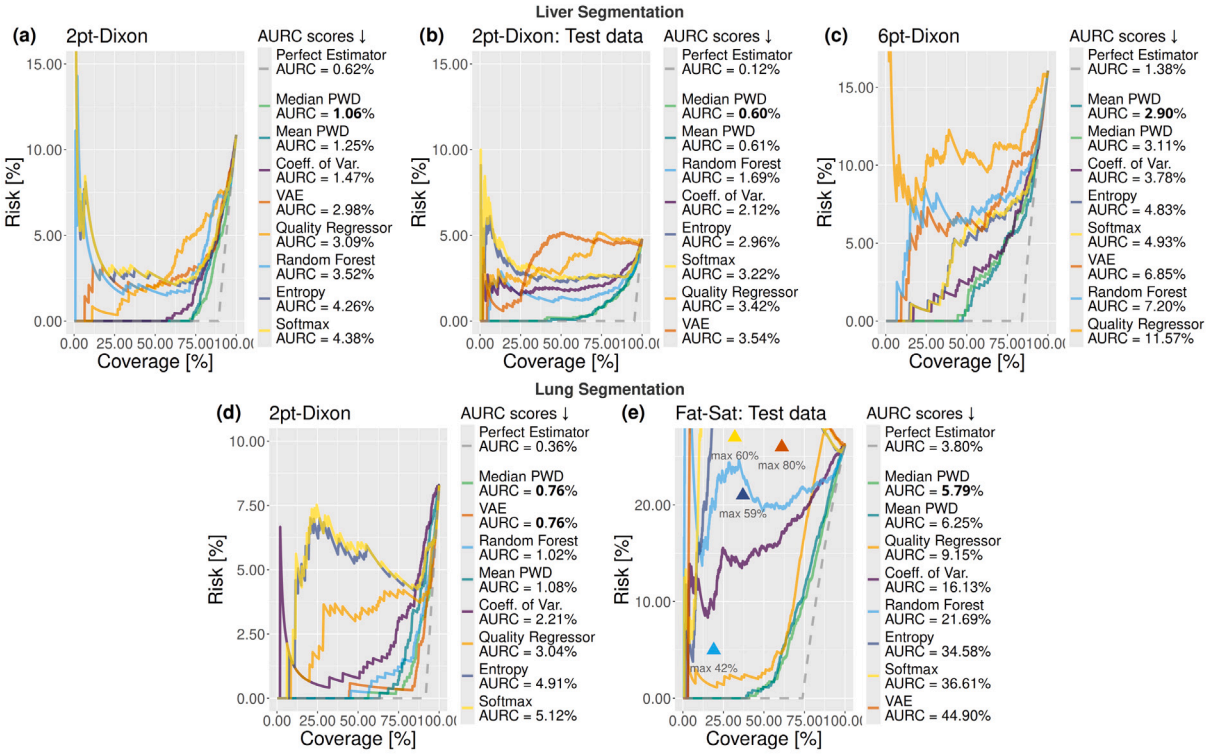


Fig. 4. Effectiveness of evaluated error estimators in performing segmentation selection: PWD-based estimators (green, cyan) showed strong performance in minimizing the risk of error inclusion (y-axis) in selected segmentations across a broad dataset coverage (x-axis). They generally outperformed softmax-, entropy-, coefficient of variation-, CNN-, VAE-, and RF regression-based error estimators (yellow, blue, purple, orange, red, and light blue) in terms of the AURC, where lower values indicate better performance. This strong performance was also observed on the held-out test datasets (b, e). Overall, the PWD-based estimators most closely matched the performance of a perfect oracle (dashed gray curve).

COSYCONET_EXP$_{\text{test}}$ ($r = -0.31$). In summary, learned estimators could perform strongly in individual settings, but were less robust across ROIs and datasets than PWD-based estimators.

Finally, the aggregation strategy influenced PWD performance. Median PWD outperformed mean PWD in all experiments except NAKO_EXP$_{\text{6pd_liver}}$, including NAKO_EXP$_{\text{2pd_lung}}$ ($r = 0.70$ vs. $r = 0.51$), COSYCONET_EXP$_{\text{test}}$ ($r = 0.61$ vs. $r = 0.38$), NAKO_EXP$_{\text{2pd_liver}}$ ($r = 0.76$ vs. $r = 0.71$), and NAKO_EXP$_{\text{2pd_test}}$ ($r = 0.59$ vs. $r = 0.55$). In contrast, on NAKO_EXP$_{\text{6pd_liver}}$, mean PWD achieved the stronger correlation ($r = 0.79$ vs. $r = 0.76$).

6.2. Selection of accurate segmentations

To evaluate the ability to select accurate segmentations in the absence of ground truth, we calculated RC-curves and AURC-scores (Section 4.2) on our error estimation datasets for the compared error estimators. As a reference point for the ideally achievable selection performance, we included a perfect oracle error estimator in the RC-curve analyses. Its uncertainty output was defined as the sequence of true Jaccard errors of the underlying predictions ($e(y, y_{\text{gt}})$ in Eqs. (7) and (10)), ordered in decreasing magnitude.

Both threshold-dependent RC-curves, based on a binary error definition, and continuous, threshold-agnostic RC-curves were considered (Section 4.2). In addition, to assess the robustness of the segmentation selection results with respect to the choice of Jaccard-based failure thresholds (τ in Eq. (7)), we performed a sensitivity analysis by evaluating the AURC of all error estimators across a range of thresholds (Appendix Fig. C.3). Figs. 4 and 5 present the resulting RC-curves and AURC scores for the NAKO and COSYCONET datasets, with the oracle estimator shown as reference (dashed gray curve).

For the threshold-dependent RC-curves, the PWD-based estimators overall showed the strongest segmentation selection performance and most closely approximated the oracle estimator. In all liver datasets and on COSYCONET_EXP$_{\text{test}}$, they achieved the lowest AURC scores (1.06%, 0.60%, 2.90%, and 5.79%, respectively), compared with the strongest non-PWD estimators in the same settings (1.25%, 1.69%, 3.78%, and 9.15%, respectively), thus indicating the lowest overall risk when selecting segmentations across increasing dataset coverage. On NAKO_EXP$_{\text{2pd_lung}}$, the median PWD and VAE jointly achieved the best AURC score (0.76%). However, their RC-curve shapes differed (see Fig. 4d): the VAE showed an earlier increase in risk at lower coverages, but subsequently reached a higher coverage at low risk than PWD. In contrast, the PWD curve avoided this early risk increase and enabled a lower-risk operating region that was no longer attainable by the VAE after its early rise. On COSYCONET_EXP$_{\text{test}}$, the quality regressor was the only non-PWD estimator with comparatively strong threshold-dependent selection performance, but it still remained clearly below the median PWD.

The continuous RC curves provided an explanatory complement to the threshold-dependent analysis, showing how well each estimator selected segmentations with low delineation error across increasing coverage, independent of a fixed failure threshold. In this analysis, PWD again performed best across all datasets except NAKO_EXP$_{\text{2pd_lung}}$, where the RF achieved the lowest continuous AURC score (1.89%), followed by the median PWD (2.00%). However, the threshold-dependent RC-curve showed that the RF placed more substantial segmentation failures earlier in the ranking than the PWD and VAE, explaining why it fell behind when failures were defined using the Jaccard threshold. A similar pattern was observed on NAKO_EXP$_{\text{6pd_liver}}$, where the RF showed strong continuous selection behavior at early, low coverages, but subsequently performed substantially worse than the PWD-based measures, a pattern that also became apparent in the threshold-dependent RC curves. These differences highlight that continuous AURC captures low-error inclusion, whereas threshold-dependent AURC captures the ability to avoid analytically relevant failure cases.

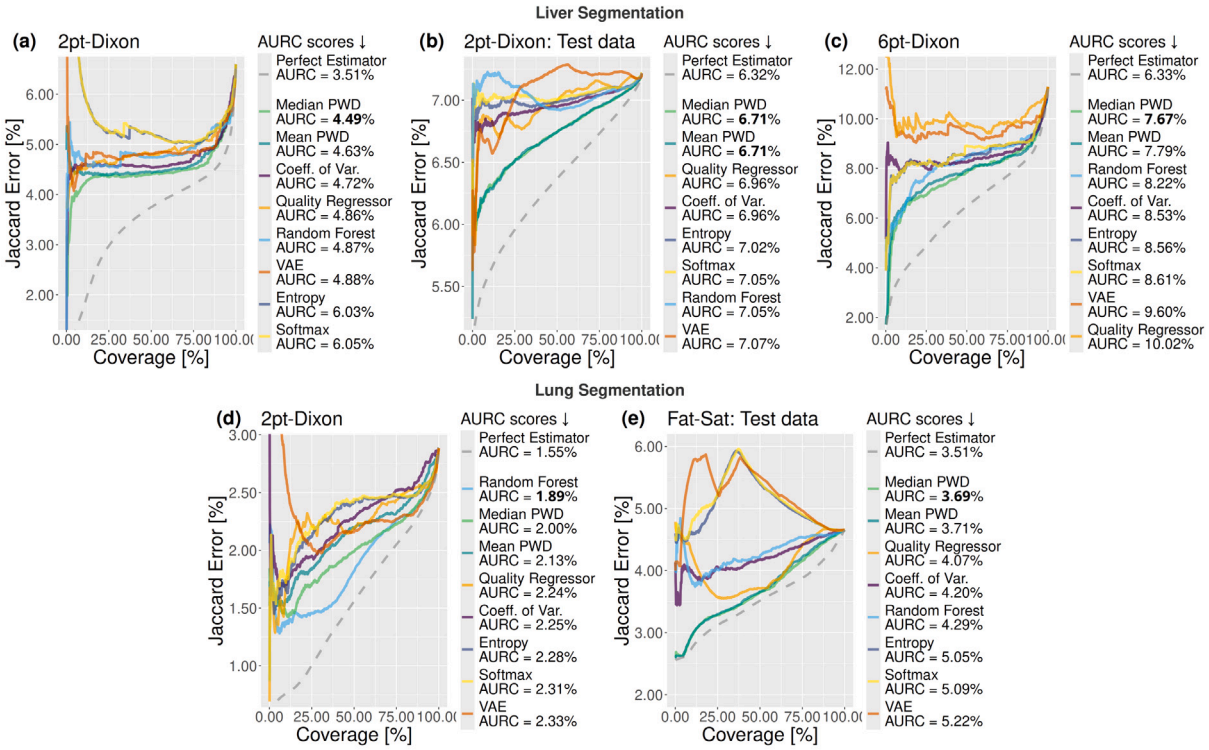


Fig. 5. Effectiveness of evaluated error estimators in performing segmentation selection: The PWD-based estimators (green, cyan) performed well in selecting segmentations that exhibit low delineation error (y-Axis: average Jaccard error) across increasing dataset coverage (x-Axis). They generally outperformed softmax-, entropy-, coefficient of variation-, CNN-, VAE-, and RF regression-based error estimators (yellow, blue, purple, orange, ochre, and light blue) in terms of the area under the curve, where lower values indicate better performance. This strong performance was also observed on held-out test datasets (b, e). Overall, the PWD-based estimators most closely matched the performance of a perfect oracle (dashed gray curve).

Participant-level bootstrap analysis further supported the threshold-dependent and continuous segmentation selection results; detailed AURC-based paired performance differences and 95% confidence intervals are reported in Appendix Fig. C.1 and Tab. C.1.

Across both RC-curve formulations, the PWD-based estimators overall provided the closest approximation to the oracle selection behavior. Their curves most consistently combined delayed risk increases, low attainable risk, and high coverage, whereas most competing estimators showed earlier risk elevations due to silent failures. This was especially relevant in the threshold-dependent analysis, where early inclusion of even a small number of failure cases strongly affected the achievable risk-coverage tradeoff.

The threshold sensitivity analysis further supported the robustness of these findings across the evaluated Jaccard-based failure thresholds (Appendix Fig. C.3).

6.3. Sampling of failure cases

To evaluate the effectiveness of the compared error estimators in sampling segmentation failures while minimizing the inclusion of error-free cases, we calculated precision–recall curves and AUPRC_F scores () on our error estimation datasets. Fig. 6 presents the PR-curves and AUPRC_F scores for the NAKO and COSYCONET datasets.

Overall, PWD-based estimators achieved the strongest failure-sampling performance in terms of AUPRC_F on all datasets except $\text{NAKO_EXP}_{\text{2pd_lung}}$. They achieved the highest AUPRC_F scores on $\text{NAKO_EXP}_{\text{2pd_liver}}$, $\text{NAKO_EXP}_{\text{2pd_test}}$, $\text{NAKO_EXP}_{\text{6pd_liver}}$, and $\text{COSYCONET_EXP}_{\text{test}}$ (75.37%, 39.83%, 75.51%, and 76.97%, respectively), exceeding the strongest non-PWD estimators in the same settings (58.96%, 30.52%, 69.14%, and 47.25%, respectively). On $\text{NAKO_EXP}_{\text{2pd_lung}}$, the VAE achieved the highest AUPRC_F (74.00%), followed by the median PWD (63.57%) and RF (59.89%). Thus, PWD-based estimators were the most robust overall, while the VAE showed the strongest performance in this individual lung setting. On $\text{NAKO_EXP}_{\text{6pd_liver}}$, mean PWD outperformed median PWD. Participant-level bootstrap analysis further supported the failure sampling results; detailed paired performance differences and 95% confidence intervals are reported in Appendix Fig. C.1 and Tab. C.1.

The learned estimators showed less consistent behavior across datasets. The VAE achieved strong failure-sampling performance on $\text{NAKO_EXP}_{\text{2pd_lung}}$, but its performance did not generalize consistently to the remaining datasets. This was particularly visible on the test datasets, where learned estimators exhibited pronounced early declines in the PR-curves. The VAE was generally preferable to the quality regressor across the NAKO datasets, but this pattern reversed on $\text{COSYCONET_EXP}_{\text{test}}$, where the quality regressor clearly outperformed the VAE (47.25% vs. 16.63%).

For the stricter task of sampling failures at 100.00% precision, several estimators achieved useful zero-false-positive sampling behavior, and the most suitable choice depended more strongly on the ROI and dataset. Among the most notable cases, median PWD was effective on the 2pt-Dixon liver data, VAE and coefficient of variation were effective on 6pt-Dixon liver data, and the quality regressor and softmax-based measure were effective on the 2pt-Dixon lung data. This indicates that zero-false-positive failure sampling is more dataset-specific than overall $\text{AUPRC_}F$ performance.

6.4. Choosing segmentation selection thresholds

To evaluate the proposed threshold search, we defined increasing risk levels in the experimental error-estimation datasets and automatically determined optimal selection thresholds using binary search and

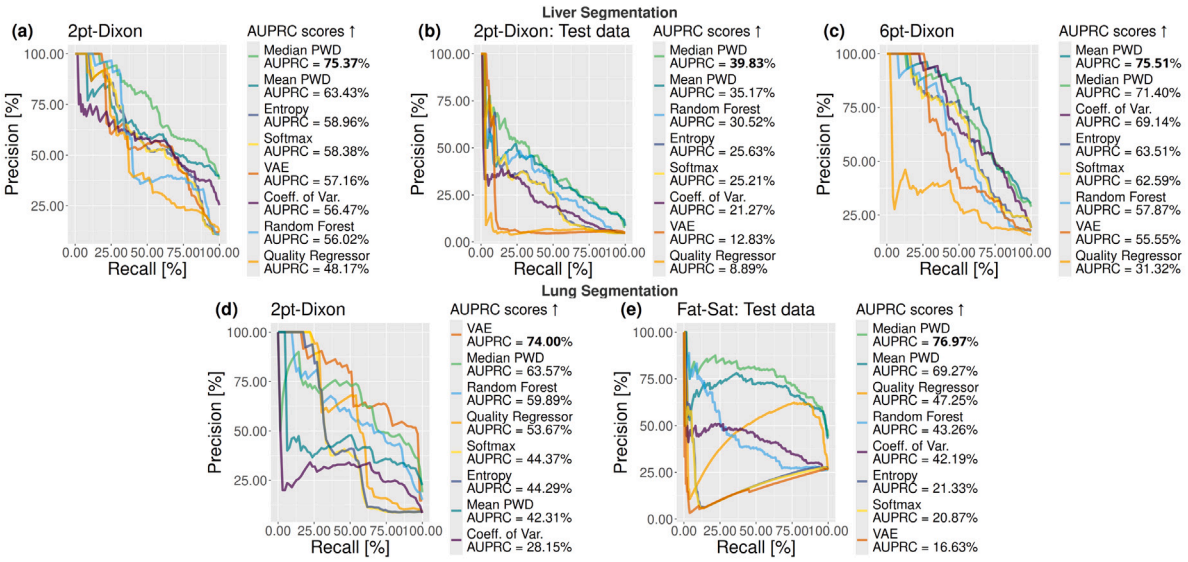


Fig. 6. Effectiveness of evaluated error estimators in sampling segmentation failures: Across liver (top row) and lung (bottom row) experiments, the PWD-based error estimators generally performed best in sampling segmentation failures in terms of the $AUPRC_F$ (higher values indicate better performance), and were only outperformed by the VAE on one dataset (d). The median PWD outperformed the mean PWD on 4 of the 5 datasets.

Table 2
Tuned operating points for segmentation quality control in the NAKO and COSYCONET cohorts: Uncertainty thresholds were automatically determined to maximize the number of selected segmentations while ensuring that upper error risk bounds met predefined target risk levels. The resulting bounds closely aligned with the target levels and retained substantial cohort coverage with a limited maximum number of expected residual errors. Lower risk bounds could not be established in the corresponding settings.

Dataset / sequence	ROI	Target risk level [%]	Upper error risk bound [%]	Training risk [%]	Test risk [%]	Cohort coverage [%]	Max expected errors out of QC-selected cases	Selection threshold
NAKO 2pt-Dixon	Liver	1.00	0.92	0.00	0.22	98.37	101/11 008	0.0105
NAKO 6pt-Dixon	Liver	2.50	2.44	0.00	–	97.32	266/10 890	0.0238
NAKO 2pt-Dixon	Lung	1.60	1.59	0.00	0.00	99.48	177/11 132	0.0076
COSYCONET FatSat	Lung	2.00	1.95	0.00	–	80.71	8/426	0.0087

risk calculation (Section 4.3). The selected operating points for downstream cohort application are summarized in Table 2, while the complete threshold-search results are provided in Appendix Tabs. C.5–C.8.

Wide zero-risk ranges occurred across the evaluated datasets, indicating that many segmentations could be selected before the first observed failure. However, within such ranges, reducing coverage does not reduce empirical risk further; instead, because the same zero empirical risk is estimated from fewer selected cases, the corresponding upper risk bound increases. To enable successful binary search convergence, we initialized the lower search bound at the end of the zero-risk range.

Across ROIs and datasets, the proposed search identified thresholds whose upper risk bounds closely matched the desired target risk levels, while empirical training risks were consistently lower. On the additional NAKO 2pt-Dixon liver test dataset, observed test risks remained well within the calculated risk bounds, supporting valid risk control under comparable NAKO test conditions. In 6pt-Dixon liver data, valid thresholds could only be established for target risks of at least 2.50%, indicating that very strict low-risk selection was more difficult in this setting.

For lung segmentation, the strictest threshold selected on NAKO also yielded low risk on $COSYCONET_EXP_{test}$, whereas more lenient thresholds produced substantially higher test risk despite increased coverage, indicating that risk levels did not transfer uniformly to this distinct target domain. COSYCONET-specific tuning yielded more finely calibrated thresholds, enabling effective QC in the more challenging COSYCONET setting.

At cohort level, selected thresholds yielded higher coverage than in the corresponding error-estimation datasets, consistent with the latter being enriched for difficult or failure-prone cases. COSYCONET coverage remained lower overall, reflecting the more challenging external domain, but COSYCONET-specific tuning still yielded meaningful coverage, despite the availability of only 13 annotated COSYCONET cases.

In summary, the proposed threshold search provided practical, risk-calibrated operating points for segmentation selection and made the risk–coverage tradeoff explicit. The results show that low-risk thresholds can be chosen reliably under comparable data conditions, whereas threshold transfer to distinct target domains requires conservative operation or domain-specific calibration.

6.5. Application in the German National Cohort

We applied the top-performing error estimators from Section 6.2 to perform segmentation QC across the entire NAKO dataset, i.e., median PWD for 2pt-Dixon liver and lung segmentations, and mean PWD for 6pt-Dixon liver segmentations. The strictest available risk-calibrated operating points were used for each setting (Table 2). Figs. 7 and 10 show error estimate distributions, applied thresholds, and resulting dataset coverages.

To assess the effect of QC on downstream measurements, we analyzed paired 2pt- and 6pt-Dixon liver volume and fat measurements before and after segmentation selection (Fig. 8). QC rejected $n = 446$ segmentations containing numerous errors associated with visible pathologies, low imaging quality, artifacts, depictions of incorrect anatomical regions, truncated livers between image blocks, or extreme liver fat and iron contents (Fig. 9). These errors had distorted the expected linear alignment between the paired measurements; their

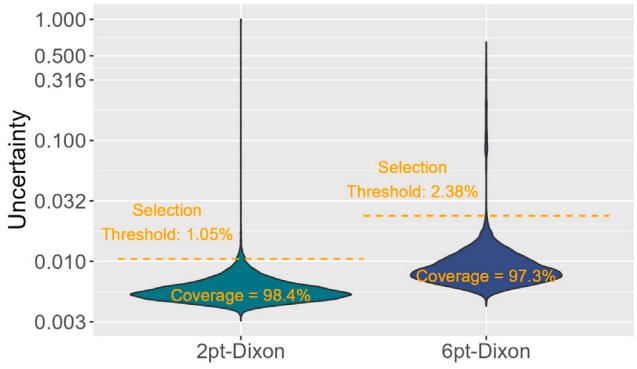


Fig. 7. Final selection of quality-controlled liver segmentations in NAKO (n = 11,190): Reliable, accurate segmentations (areas below the dotted lines, covering > 97.00% of the dataset) were selected in 2pt-Dixon (left) and 6pt-Dixon data (right), while numerous errors and abnormalities (areas above the dotted lines) were filtered out.

removal effectively restored this relationship, increasing Pearson correlations from $r = 0.97$ to $r = 0.99$ for liver volume and from $r = 0.36$ to $r = 0.98$ for liver fat ($p < 0.01$, Steiger's test). Corresponding bootstrap estimates are reported in Appendix C.2.

Many rejected segmentation errors and abnormalities would not have been apparent from volume or fat measurements alone (Fig. 8). A few outliers ($n < 15$) were not flagged during QC. These were attributed to swap artifacts, truncated livers, and extreme liver fat and iron content, yet the trained models confidently performed segmentation on these cases.

For lung segmentation, QC identified $n = 59$ substantial failures, which were largely indistinguishable from valid cases based on lung volume alone (Fig. 11). Visual inspection revealed failures associated with several artifact types (Fig. 12), whereas extreme lung volumes attributable to plausible anatomical variation were correctly retained. Thus, QC removed subtle segmentation failures while preserving physiologically plausible population extremes.

6.6. Application in the COSYCONET cohort

For COSYCONET, we used the median PWD error estimator and adopted the strictest COSYCONET-tuned operating point determined in Section 6.4 (Table 2). The corresponding uncertainty distribution is shown in Fig. 10. Compared with NAKO, COSYCONET showed higher uncertainty and lower coverage was achieved (80.71% vs. > 99%), consistent with the more challenging nature of the data, including lower lung-to-background contrast, particularly near the diaphragm, stronger pathological abnormalities, and more frequent artifacts.

We assessed the effect of QC on paired CT- and MRI-derived lung volumes and on segmentation-mask characteristics (Fig. 11). Unlike the paired NAKO liver measurements, CT and MRI lung volumes are not expected to show perfect linear alignment because of differences in respiratory state, acquisition protocols, and image contrast. Therefore, physiologically plausible deviations between CT and MRI lung volumes were expected and should not be rejected by QC.

Segmentation failures produced the largest CT-MRI volume discrepancies and distorted the paired CT–MRI lung volume relationship, with the rejected cases showing substantially lower correlation ($r = 0.78$). QC rejected $n = 102$ of 528 cases, increasing the overall Pearson correlation from $r = 0.87$ to $r = 0.90$ ($p < 0.01$, bootstrap test; Appendix Tab. C.3) while retaining plausible respiratory differences between modalities. Many of these failures were located within the main volume distribution and would therefore likely have been missed by plausibility checks based on lung volume alone.

The excluded failure cases also showed substantial inconsistencies in their segmentation masks, including scattered mask fragments, large holes, and lower sphericity (Fig. 11C-E). Corresponding statistical analyses are reported in Appendix C.2. Visual inspection confirmed heterogeneous failure causes, including challenging imaging conditions, false image inputs, image composition artifacts, motion artifacts, metal artifacts, and pathological abnormalities (see Fig. 12).

7. Discussion

Our experiments demonstrate that uncertainty-based error estimation can be employed with great success for large-scale segmentation QC. The following discussion considers these findings in terms of QC performance, threshold optimization, the proposed evaluation framework, cohort-level application across NAKO and COSYCONET, generalizability, and practical efficiency.

7.1. Effective segmentation quality control using ensemble uncertainty

Our results demonstrate that the proposed use of the PWD error estimator enables effective uncertainty-based segmentation QC for large imaging datasets, particularly by selecting segmentations at low error risk while maintaining broad dataset coverage.

The absence of silent failures was a prerequisite for successful selection. Such failures were observed for most compared, non-PWD estimators. An exception was observed for the VAE and RF in the $\text{NAKO_EXP}_{\text{2pd_lung}}$ dataset, where these methods also showed low error risk across a wide range of coverages. However, this behavior was not consistent across datasets, particularly during test set application.

The PWD error estimator also demonstrated a notable alignment with segmentation delineation accuracy across datasets, modalities and target organs. Among the compared estimators, learning-based approaches also achieved strong correlations, but only in selected settings.

Finally, the PWD enabled effective sampling of error and outlier cases from the high-uncertainty range. Within this range, a significant proportion of all error cases in the dataset was identified with high precision, which proved highly useful for error inspection across the NAKO and COSYCONET datasets. Learning-based approaches also performed well in selected settings and may therefore be considered depending on ROI, dataset, and specific QC objective. The observed failure sampling performances hold promise for training-candidate selection in active learning (Settles, 2009; Beluch et al., 2018). On 6pt-Dixon data, the coefficient of variation may complement the PWD when minimizing false positives is critical.

Our results align with previous smaller studies (Roy et al., 2018; Hoebel et al., 2020), where the PWD also outperformed other uncertainty measures in reflecting delineation accuracy. However, PWD showed generally weaker correlations with segmentation accuracy than reported previously for the dedicated accuracy estimators from Robinson et al. (2017, 2019, 2018), Wang et al. (2020), Hann et al. (2021) and the uncertainty-based approaches from Roy et al. (2018), Hoebel et al. (2020). The former difference is expected, as these methods specifically predict delineation accuracy, while the latter may partly reflect the substantially larger-scale evaluation performed here. Learning-based estimators showed less robust performance under the low training-budget conditions and domain shifts investigated in this study. Similar dataset-dependent correlation strengths of the PWD have been reported in subsequent comparisons across other small datasets (Zenk et al., 2025; Lennartz and Schultz, 2025). More broadly, large-scale utility of uncertainty quantification has also been demonstrated for out-of-distribution (OOD) detection in digital pathology (Linmans et al., 2022).

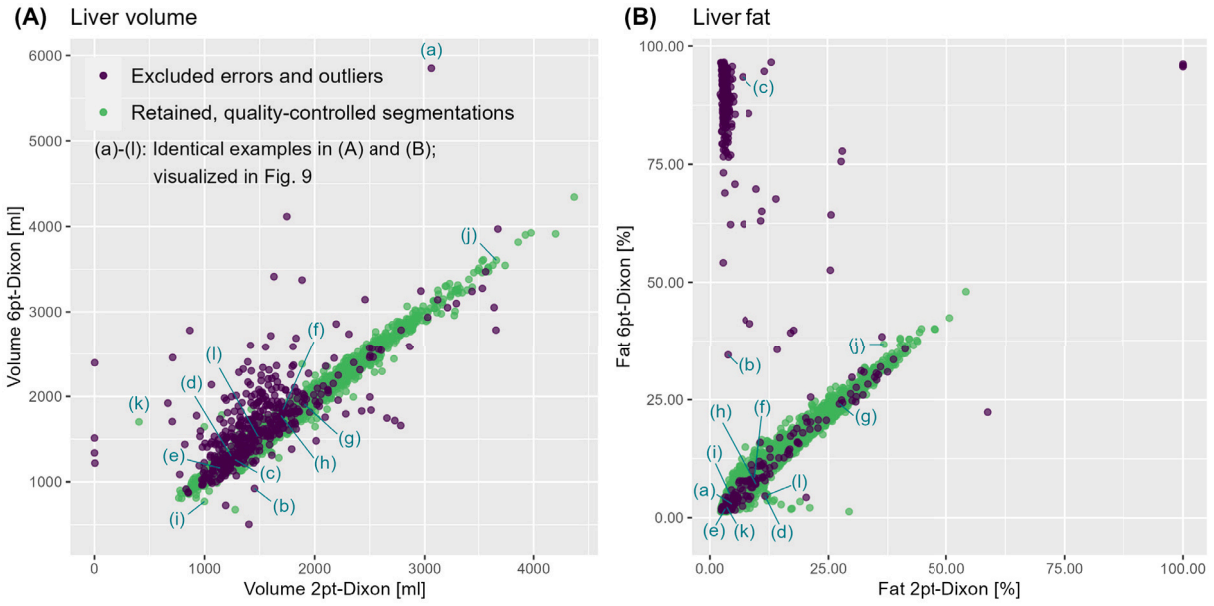


Fig. 8. Liver volume and fat measurements before and after segmentation quality control (QC): Paired volume (A) and fat (B) measurements from the segmented livers in 2pt- and 6pt-Dixon data should align along the *x*- and *y*-axes, forming a linear relationship between the two sequences. This alignment was only observed for the selected, quality-controlled segmentations (green points). The selection process effectively excluded most measurement outliers (violet points), restoring the validity of the measurements. The Pearson correlation coefficient increased for volumes from $r = 0.97$ to $r = 0.99$, and for fat from $r = 0.36$ to $r = 0.98$ ($p < 0.01$, Steiger's test). Few outliers remained, due to confident segmentations performed under challenging imaging conditions, including swap artifacts, truncated livers between image blocks, and extreme liver fat or iron content. Various errors were identified in the rejected segmentations, which would have gone unnoticed if QC had been solely based on volume and fat measurements (violet points within the green point clouds). Points labeled (a)–(l) correspond to example cases shown in Fig. 9a–l.

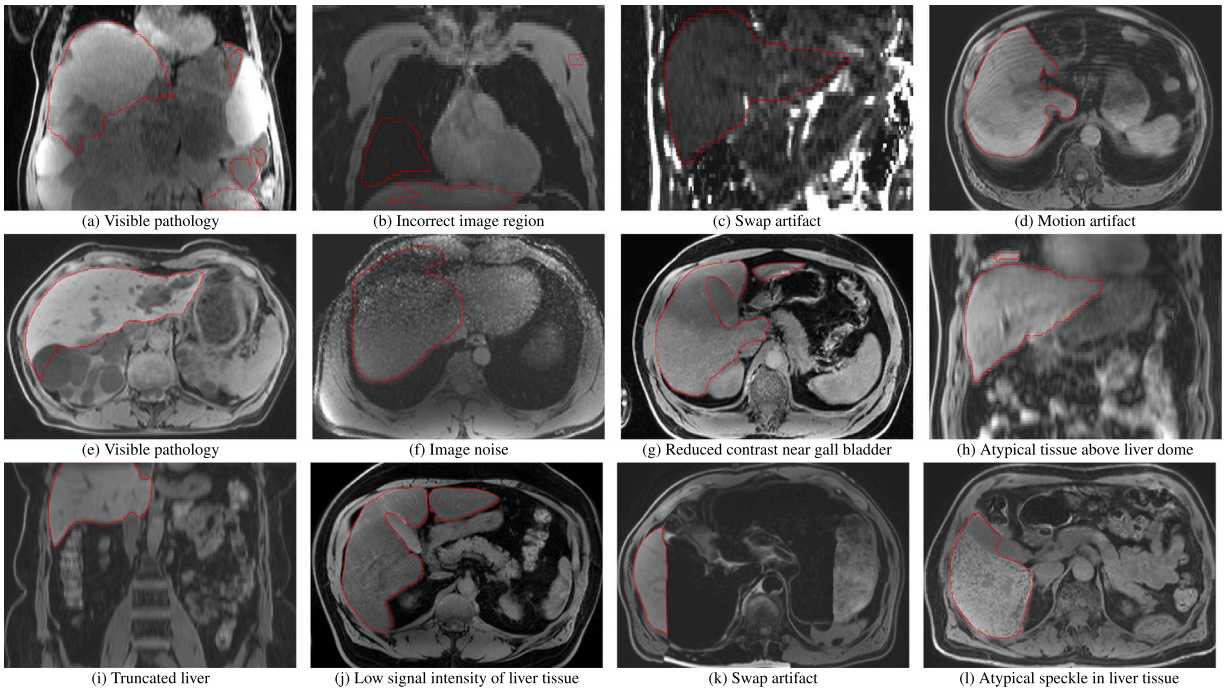


Fig. 9. Visual examples of conspicuous segmentations: The proposed quality control effectively identified a substantial number of erroneous and abnormal cases (examples: a–h). Most of these aberrations were linked to visible pathologies (a, e), incorrect or misaligned image regions (b, i), artifacts and poor image quality (c, d, f, h, k, l), or extreme liver fat and iron content (j, l). A small number of abnormal cases were not excluded during segmentation selection. In these instances, confident segmentations were still generated despite underlying image abnormalities (i–l). The cases in (i, k) contain unambiguous segmentation targets that remain valid in the presence of the underlying abnormalities and were accurately delineated. These cases therefore represent examples that segmentation QC methods are not inherently designed to identify. Such imaging artifacts must be separately detected and addressed. In contrast, the cases in (j, l) represent valid segmentations of conspicuous imaging findings that may be relevant for downstream metabolic analyses and were appropriately accepted during QC.

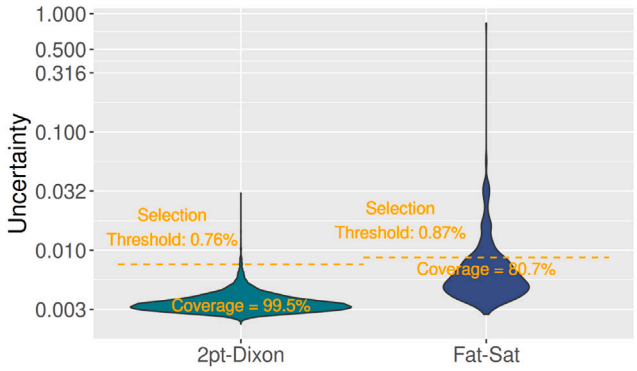


Fig. 10. Final selection of quality-controlled lung segmentations in the NAKO (n = 11,190) and COSYCONET (n = 528) datasets: Reliable, accurate segmentations (areas below the dotted lines) were selected in 2pt-Dixon data of NAKO (left, covering > 99.00% of the dataset), and in Fat-Sat data of COSYCONET (right, covering > 80.00% of the dataset). Numerous errors and abnormalities were filtered out in both datasets (areas above the dotted lines).

7.2. Automated threshold search for optimal segmentation selection

Our results demonstrate that the proposed threshold search enables an informed decision to accept or reject segmentations from imperfect error estimates, by identifying operating points that balance dataset coverage against residual error risk. In our experiments, identifying such operating points proved non-trivial due to the large number of possible thresholds and the non-monotonic RC curve. Moreover, the use of upper risk bounds to account for potential generalization errors across the full cohort further constrained threshold selection. Evaluating every possible threshold would additionally require increasingly strict correction of the residual risk δ with the number of search iterations (Section 4.3), potentially yielding overly conservative bounds.

The proposed automatic search successfully identified thresholds whose upper risk bounds closely matched the target risk levels while maximizing coverage, requiring less than 5 s per risk level in our experiments. This specifically enabled weighing error risk against attainable dataset coverage for specific downstream analyses. For instance, in the NAKO 2pt-Dixon liver data ($n = 11\,190$), fewer than 112 remaining segmentation errors were expected when rejecting the ∼180 most uncertain cases (1.00% target risk, Tab. C.5), whereas rejecting only the ∼30 most uncertain segmentations resulted in an expectation of less than 504 remaining errors (4.50% target risk level in Tab. C.5). For downstream analysis of liver fat, rejecting 180 cases seems advisable if the error risk can be limited to 112 cases. In scenarios where rejecting cases should be avoided (e.g., to maintain balanced stratification) and where higher error rate is tolerable, alternative options may be considered.

Similarly, in the NAKO 2pt-Dixon lung data ($n = 11\,190$), the strictest target risk level required rejecting fewer than 60 cases while already retaining 99.48% of the cohort (Tab. C.7). Relaxing the threshold would retain only ∼54 additional cases while increasing the maximum expected residual errors from 177 to 210. Given that the rejected cases included substantial segmentation failures (Fig. 12), the stricter threshold represented a more favorable tradeoff in this setting.

Previous work on large-scale segmentation QC has not explicitly addressed optimization of the RC tradeoff (Section 2.3). This gap may have limited the development of more general QC approaches capable of reliably selecting high-quality segmentations for downstream analyses. Notably, the presented threshold optimization is not specific to the PWD and can be applied to other error estimators used for segmentation selection.

A limitation of the presented approach is that risk calculations are based on the sampled segmentation-uncertainty pairs described in Section 5.4, which may not encompass all possible error types. As such, an elevated risk of silent failures across NAKO would have been a notable concern. In NAKO, however, the sampled error characteristics appeared sufficiently representative: no unexpected discrepancies were observed in the additional 2pt-Dixon liver test dataset or the final cohort analysis, and silent failures were solely encountered when confident segmentations were produced under abnormal circumstances (see Fig. 9i–l). The NAKO lung application further supported reliable operating point selection, with the selected threshold identifying otherwise inconspicuous failures while preserving plausible anatomical and physiological extremes.

The COSYCONET experiments, however, demonstrate that calibration data representativeness becomes more critical under external domain shift. Although the strictest NAKO-derived threshold retained low empirical risk in COSYCONET, less conservative operating points did not preserve the same risk levels. The tolerated failures were qualitatively similar across datasets but substantially more prevalent in COSYCONET, including challenging lung boundaries near the diaphragm. The same consistency of failure characteristics across datasets was also observed for liver segmentation, where residual errors mainly involved boundary ambiguities near adjacent structures such as the intestines or diaphragm. In contrast to the transferred NAKO-derived thresholds, direct calibration on COSYCONET provided interpretable risk-coverage tradeoffs. This suggests that the proposed upper risk bounds are most meaningful when calibration samples sufficiently represent the target distribution; under anticipated domain shift, separate calibration, additional representative samples, or more conservative operating points may be required.

In addition, the upper risk bound, designed to conservatively account for risk on unseen samples within the underlying cohort, showed reliable behavior across all examined 2pt- and 6pt-Dixon datasets. Future work could further improve calibration by incorporating additional candidate cases through anomaly detection and by explicitly considering characteristics of the underlying data distribution.

7.3. Detailed evaluation of segmentation quality control

Our findings confirm the value of task-specific performance measures for segmentation QC. The RC-curve effectively characterized segmentation selection performance by revealing susceptibility to silent failures, attainable low-risk regions, and achievable dataset coverage. The AURC summarized these characteristics for overall comparison of error estimators. Similarly, PR-curves and AUPRC characterized failure sampling performance.

The continuous and threshold-dependent RC-curves differed in specific cases around the chosen failure threshold. Non-zero errors could cause early increases in the continuous curve without constituting failures, whereas small changes around the threshold could markedly affect the threshold-dependent curve. Visual inspection confirmed that these cases represented distinct silent failures rather than artifacts of the Jaccard-based error definition (Appendix C.6 and Fig. C.4). This behavior was observed for the RF in the NAKO_EXP$_{6pd_liver}$ and the VAE in NAKO_EXP$_{2pd_lung}$ datasets. Joint interpretation of both variants therefore provided a more complete characterization of segmentation selection performance.

The oracle curves proved useful in supporting the heuristic choice of Jaccard-based failure thresholds. The selected thresholds remained close to the highest values for which the oracle still indicated meaningful selection, suggesting that this criterion may help guide threshold selection in future applications.

Most studies on segmentation QC offer limited insight into the effectiveness of error estimators for dataset-level segmentation QC (Section 2.2), particularly for segmentation selection. To our knowledge, only Ng et al. (2022) introduced a tailored performance measure to evaluate a specific dataset-level aspect of segmentation QC.

The need for additional, goal-specific evaluation measures has also been highlighted by Linmans et al. (2022) in the context of large-scale

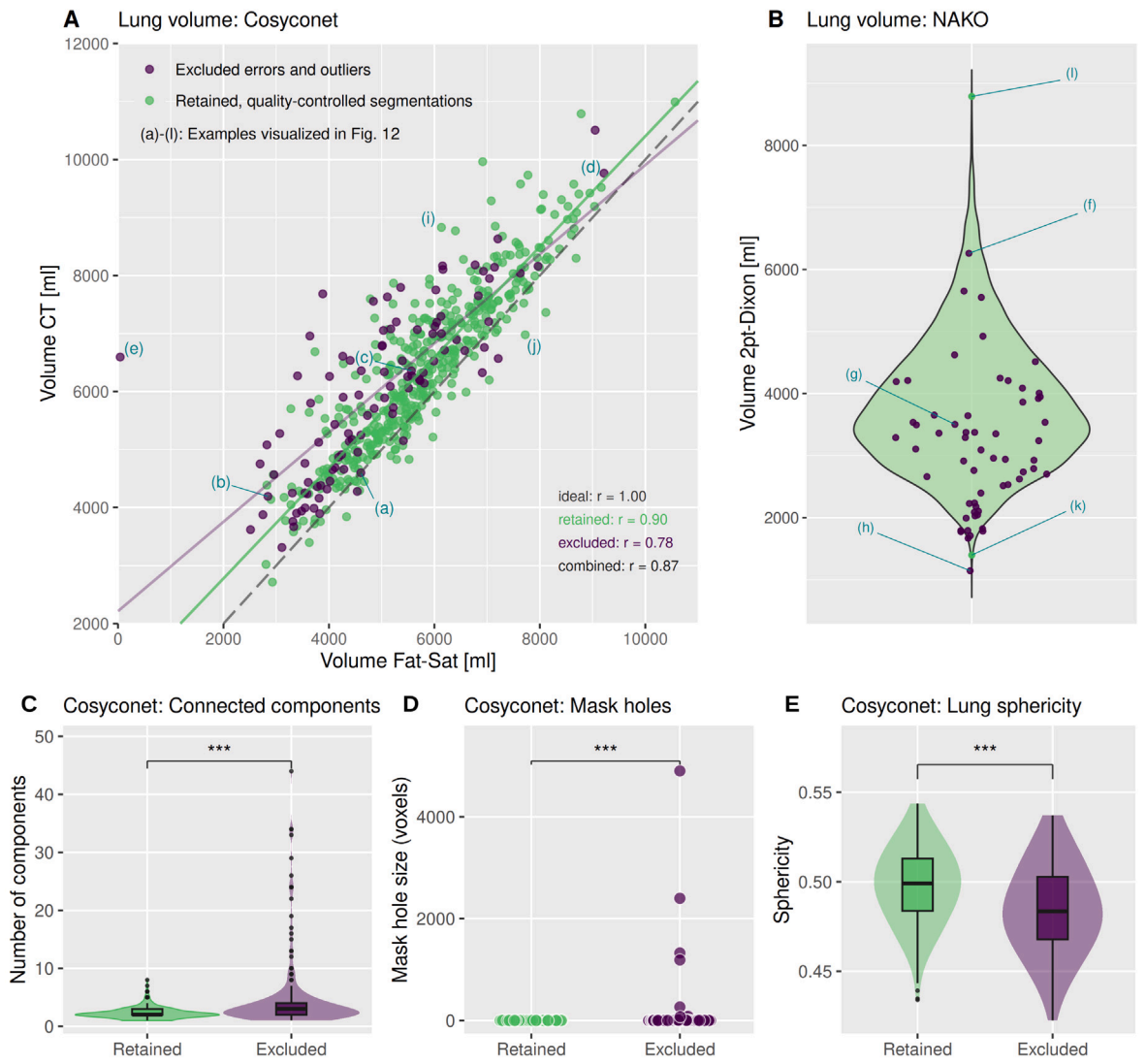


Fig. 11. Lung volume measurements and segmentation mask characteristics before and after segmentation quality control (QC): Paired volume measurements are shown for MRI-Fat-Sat and CT lung segmentations in COSYCONET (x and y-axes in A), and from 2pt-Dixon lung segmentations in NAKO (B). Ideally, the paired CT–MRI measurements would follow a linear relationship (dashed line in A), with a bias toward larger segmented lung volumes in CT data (shift in the y-direction). The proposed QC excluded numerous segmentation failures (violet points), while retaining reliable quality-controlled segmentations (green points and shaded areas). The identified failure cases exhibited substantial distortion in the paired CT–MRI lung volume relationship ($r = 0.78$, violet line in A). This distortion was mitigated by the proposed QC, increasing the Pearson correlation coefficient across all paired measurements from $r = 0.87$ before QC to $r = 0.90$ after QC (green line in A; $p < 0.01$, bootstrap test), thereby approximating the expected linear relationship. The excluded failure cases also exhibited substantial inconsistencies in their segmentation masks (purple plots in C–E), including scattered mask fragments (C), large holes within the segmented lung regions (D), and an altered lung mask sphericity (E). These inconsistencies were effectively removed from the quality-controlled segmentations (green distributions in C–E, *** indicates $p < 0.001$, two-sided Wilcoxon rank-sum test). Various errors would have been missed if only a manual outlier inspection of volume measurements or mask characteristics had been performed (e.g., violet points within the green point clouds/plot areas of A, B). In COSYCONET (A), the largest discrepancies between MRI- and CT-derived lung volumes attributable to segmentation failures were eliminated, while accurate segmentations, capturing respiratory differences between MRI and CT acquisitions, were retained (e.g., cases i and j in A; see Fig. 12 i,j). Errors were caused by challenging imaging conditions, false image inputs, image composition artifacts and pathological abnormalities (Fig. 12). Points labeled (a)–(l) correspond to example cases shown in Fig. 12a–l.

OOD detection. For future investigations, it may be valuable to assess how the chosen segmentation quality criterion affects error estimation. Methods that explicitly incorporate volumetric overlap measures, including the CNN-, VAE-, or RF-based Dice regression methods as well as the PWD-based estimators, may be particularly well aligned with overlap-based quality criteria. Conversely, entropy-, softmax-, or coefficient-of-variation-based measures may be less directly aligned with such criteria. Future work could explore error estimators that explicitly account for other segmentation performance measures where these are more relevant.

7.4. Application to large-scale imaging data

Our results highlighted that segmentation errors can affect downstream imaging analyses both globally and locally. In NAKO, errors distorted the overall relationships between paired 2pt- and 6pt-Dixon measurements, while individual errors could also shift measurements across clinically relevant decision thresholds (Fig. 1). Such errors occur in practice with fully optimized nnU-Net ensembles in challenging cases. Similar effects may arise from lung segmentation errors in COSYCONET through distorted volume or shape measurements. This is particularly relevant near the diaphragm or lung apex, where shape

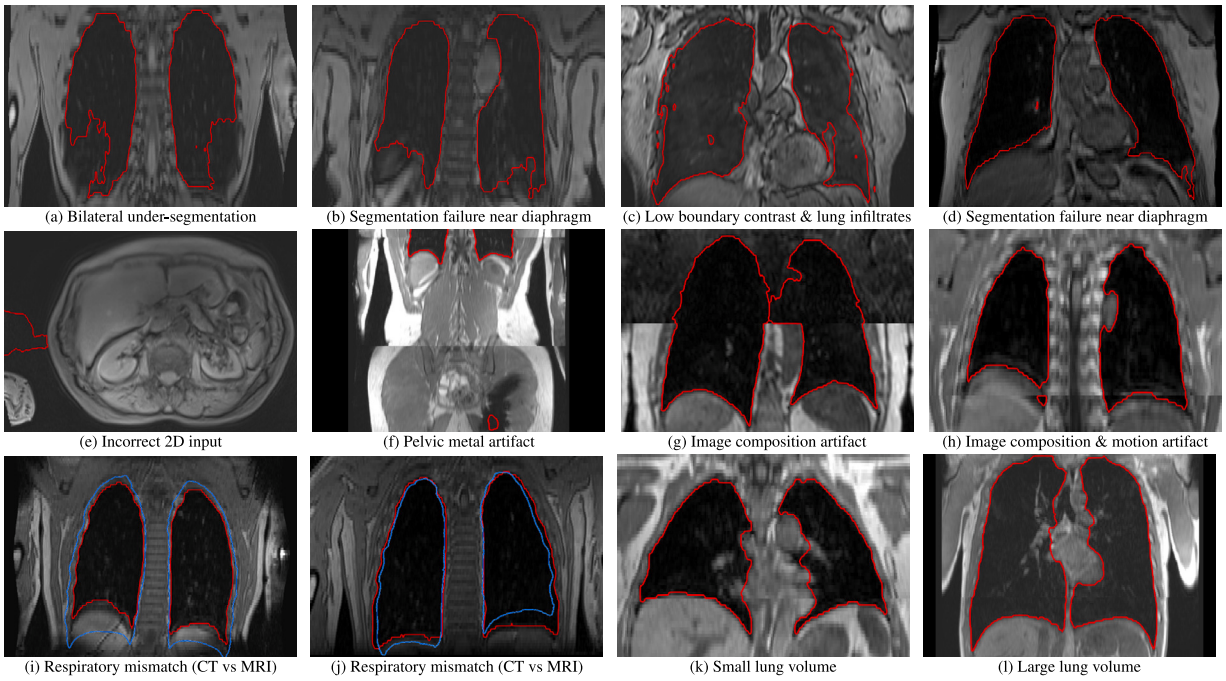


Fig. 12. Visual examples of conspicuous segmentations: The proposed quality control (QC) effectively identified a wide range of segmentation errors (examples: a–h), arising from challenging imaging conditions (a–d), visible pathologies (c, d), incorrect image input (e), and imaging artifacts (f–h). In contrast, accurate segmentations of cases with extreme lung measurements were retained during QC, including those with respiratory mismatches between MRI- and CT-acquisitions (i, j), as well as particularly small or large lung volumes (k, l).

information may be informative in diseases such as COPD. Such errors can be particularly consequential in stratified subgroup analyses, where smaller sample sizes and restricted value ranges can amplify their impact.

Our results underline the practical advantage of automated QC at population scale. Manual identification of the aforementioned failure cases could require hundreds or thousands of visual inspections across the full range of volume or fat measurements (e.g., for identifying the violet cases in Figs. 8 and 11). Additionally, paired acquisitions are not always available to facilitate plausibility checks. Moreover, the proposed QC removed ill-shaped and inconsistent segmentation masks that are unsuitable for downstream analyses (Figs. 11C–E, and 12). Lastly, analyzing the rejected results offered valuable insights into recurring segmentation errors and dataset abnormalities (Figs. 9 and 12).

Our findings confirm and extend previous research into the importance of segmentation QC in large population-based datasets (Kart et al., 2022). In line with our study, segmentation errors were found to affect geometric measurements and to be associated with imaging quality and truncation effects. Beyond these findings, we observed effects on imaging-derived fat measurements and additional error causes, including visible pathologies, misrepresented image regions, and extreme liver contents. Kart et al. (2022) also emphasized the need for automated QC to alleviate extensive manual workload. Furthermore, the proposed method effectively detected less obvious delineation errors that were reported to be challenging to identify through the visual QC approach employed by Kart et al. (2022).

One limitation of the proposed method is the relatively high effort required for segmentation model training to generate the necessary segmentation–uncertainty samples (Table 1). Exploring alternative strategies to diversify segmentation model outputs - such as applying random dropout masking during inference, introducing stochastic image quality degradation (Ng et al., 2022), or leveraging partially trained models from earlier stages of training (Yazdan-Panah et al., 2024) - could be a valuable direction for future work. In our study, model pre-training and iterative refinement were employed to enable reliable liver segmentation despite challenging variability in liver shape and pathology. This effort may be lower for other ROIs or imaging conditions, as illustrated by the presented NAKO lung segmentation scenario.

7.5. Method generalizability across datasets and tasks

Extending the proposed method beyond the segmentation tasks, target organs, and imaging settings investigated in this study remains a meaningful direction for future work, including smaller anatomical structures and ROIs with different shapes. The method was primarily developed for challenges that are particularly relevant at large scale, including impractical manual QC, runtime constraints, and limited annotated training data across the many ROIs investigated in population imaging cohorts. Nevertheless, we expect the method to be beneficial across a broader range of dataset sizes, including cohorts comparable to COSYCONET, and across imaging modalities beyond MRI.

Methodologically, the proposed error estimation relies solely on structural deviations in the output segmentation masks and is therefore agnostic to the underlying imaging characteristics. This may contribute to its robustness across acquisition settings, as the PWD remained effective on test data with distinct imaging characteristics. Segmentation uncertainty frequently arose from model disagreements at specific anatomical locations, particularly where tissue contrast or parenchyma appearance was affected by imaging conditions or visible pathologies (Figs. 9 and 12). In such cases, uncertainty reflected the spatial extent of disagreement between predictions rather than the underlying imaging characteristics.

Learning-based estimators may generalize when the learned features remain informative across domains, as indicated by the quality regressor on COSYCONET. The VAE and RF also depend on previously observed segmentation-error patterns through learned mask representations or spatial uncertainty characteristics. This may explain their reduced generalization from NAKO to COSYCONET, where additional error patterns occurred.

For applications to other ROIs, representative sampling of potential failure cases will be important (Section 5.4). Such samples are needed to capture anatomy-dependent uncertainty characteristics and to calibrate PWD overlap values, which depend on target-organ size. This may be more straightforward for difficult segmentation tasks, where diverse failures occur more naturally. For datasets with a high prevalence of pathology, consistent ground-truth definitions and pathological examples will additionally be important, as illustrated by the COSYCONET results. Ideally, such annotations are already available from segmentation model training.

7.6. Runtime and resource analysis

To assess computational and data efficiency for large-scale segmentation QC, we compared the proposed method with established QC approaches under the same low-data regime and hardware conditions. The comparison included methods previously applied at scale with substantially larger labeled training sets. Experiments were conducted on a workstation equipped with an AMD Ryzen 9 3900X central processing unit (12 cores), 64 GB of system random-access memory, and a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB video random-access memory).

All compared approaches achieved fast processing times which increased from the smaller abdominal 6pt-Dixon liver segmentations to the larger whole-body 2pt-Dixon lung segmentations. Runtimes ranged from 0.66–1.13, 0.65–0.87, 3.44–8.57, and 0.29–4.48 seconds per case for the VAE-, CNN-regression-, RF-, and PWD-based approaches, respectively. Setup and model loading times of the VAE-, CNN-regression-, and RF-approaches, which would not account substantially in large-scale applications, were excluded from the measurements. These measurements assume that 5-fold nnU-Net ensembles (Isensee et al., 2021) are already available as the default segmentation solution.

If ensemble segmentations are not already available, for example when using single-model or pre-existing segmentations, ensemble-based QC requires additional inference to generate the necessary predictions. In our setup, ensemble segmentations were generated in parallel on a GPU cluster using Titan RTX (23.6 GB) and Tesla V100 PCIe (32 GB) nodes. Per-node inference times for a five-model ensemble run of nnU-Net ranged from 3.5–5.8 seconds per case for the 6pt-Dixon and 2pt-Dixon liver datasets, respectively, and reached 36.9 s per case for the whole-body lung dataset. For the ensemble-based QC approaches, these inference times would add to the overall processing time in the aforementioned scenarios. In contrast, the VAE- and CNN-based approaches would be readily applicable to underlying single segmentations.

Finally, the maximum softmax-, entropy-, and RF-based approaches require storing the softmax outputs of the ensemble segmentations, adding 12.0–45.4 MB per ensemble segmentation in our experiments.

In summary, the proposed method enables fast and reliable segmentation QC across a wide range of scenarios, making it an effective and widely available solution for large-scale applications.

8. Conclusion

In this work, we presented an efficient uncertainty-based approach for segmentation QC in large population-based imaging cohorts such as NAKO or UKBB. The proposed framework supports key QC tasks in the absence of ground truth, including dataset-level assessment of delineation accuracy, sampling of segmentation failures, and risk-calibrated selection of accurate segmentations. It further provides an automatic risk-aware threshold tuning strategy for choosing practical segmentation-selection operating points. Our experiments showed that task-specific evaluation is essential for assessing these different QC objectives, with risk–coverage analyses serving as a key component for evaluating segmentation selection. Error estimation based on the PWD of ensemble outputs showed the most consistent strong performance across tasks, datasets, and organs compared with the other evaluated estimators, including learning-based approaches.

Applied to more than 10,000 automatically segmented cases in NAKO and to the external COSYCONET cohort, the proposed method enabled fast and effective QC by filtering numerous segmentation errors and outliers and improving the reliability of derived imaging measurements. At the same time, the PWD-based approach remained simple and efficient because it required only ensemble segmentation outputs, avoiding the additional training complexity and stronger annotation requirements of learning-based error estimators. These findings demonstrate that the proposed uncertainty-based segmentation QC can effectively replace extensive manual review efforts while supporting reliable downstream analyses in future large-scale imaging studies.

CRediT authorship contribution statement

Tobias Norajitra: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Tobias Nonnenmacher:** Writing – review & editing, Resources, Data curation, Conceptualization. **Tobias Haueise:** Writing – review & editing, Validation. **Fabian Isensee:** Writing – review & editing, Software, Resources. **Maximilian Zenk:** Writing – review & editing, Validation. **Lea J. Michel:** Writing – review & editing, Resources, Data curation. **Jan Warszawski:** Writing – review & editing, Resources, Data curation. **Marc-Nicolas von Itter:** Writing – review & editing, Validation, Project administration. **Steffen Ringhof:** Writing – review & editing, Resources, Project administration. **Thoralf Niendorf:** Writing – review & editing, Resources. **Jessica Krepel:** Writing – review & editing, Resources. **Manuela Harries:** Writing – review & editing, Resources. **Rafael Mikolajczyk:** Writing – review & editing, Resources. **Jochen Hirsch:** Writing – review & editing, Resources. **Daniel C. Hoinkiss:** Writing – review & editing, Resources. **Thomas Kroencke:** Writing – review & editing, Resources. **Josua Decker:** Writing – review & editing, Resources. **Filippo Michelotti:** Writing – review & editing, Resources. **Christoph M. Friedrich:** Writing – review & editing, Resources. **Johannes Haubold:** Writing – review & editing, Resources. **Tobias Pischon:** Writing – review & editing, Resources. **Sigrid Thierry:** Writing – review & editing, Resources. **Stefan N. Willich:** Writing – review & editing, Resources. **Carsten O. Schmidt:** Writing – review & editing, Resources. **Jakob Weiß:** Writing – review & editing, Resources, Data curation. **Susanne Rospleszcz:** Writing – review & editing, Project administration, Funding acquisition. **Hans-Ulrich Kauczor:** Writing – review & editing, Resources, Funding acquisition. **Jürgen Machann:** Writing – review & editing, Resources, Funding acquisition. **Fritz Schick:** Writing – review & editing, Resources, Funding acquisition. **Christopher L. Schlett:** Writing – review & editing, Resources, Funding acquisition. **Fabian Bamberg:** Writing – review & editing, Resources, Funding acquisition. **Johanna Nattenmüller:** Writing – review & editing, Resources, Funding acquisition. **Klaus Maier-Hein:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. All citations were manually provided.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: FB holds shares in Siemens Healthineers. He receives financial support from Siemens, Philips, Sanofi, and Bayer.

Acknowledgment

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 428224476/SPP2177. This project was conducted with data (Application No. NAKO-308/752) from the German National Cohort (NAKO) (www.nako.de). NAKO is funded by the Federal Ministry of Education and Research (BMBF) [project funding reference numbers: 01ER1301A/B/C, 01ER1511D 01ER1801A/B/C/D, and 01ER2301A/B/C], federal states of Germany and the Helmholtz Association, the participating universities and the institutes of the Leibniz Association. We thank all participants who took part in the NAKO study and the staff of this research initiative. This study was supported in part by a grant (01GI0925) from the Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). The COSYCONET study is supported by the German Center for Lung Research (DZL), grant number 82DZLI05A2 (COSYCONET), and the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung [BMBF]), grant number 01GI0881, and is furthermore supported by unrestricted grants from AstraZeneca GmbH, Boehringer Ingelheim Pharma GmbH & Co. KG, GlaxoSmithKline GmbH&Co. KG, Grifols Deutschland GmbH, and Novartis Deutschland GmbH.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2026.104351.

Data availability

Data from the NAKO, HELENA, and COSYCONET studies used in this work are subject to data usage agreements that prohibit their public release.

References

Ahmed, F., Courville, A., 2020. Detecting semantic anomalies. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 3154–3162.

Alipanahi, B., Hormozdiari, F., Behsaz, B., Cosentino, J., McCaw, Z.R., Schorsch, E., Sculley, D., Dorfman, E.H., Foster, P.J., Peng, L.H., 2021. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. Am. J. Hum. Genet. 108 (7), 1217–1230.

Anastasopoulos, C., Reisert, M., Kellner, E., 2017. Nora imaging: A web-based platform for medical imaging. Neuropediatrics 48, S1–S45.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Carapella, V., Kim, Y.J., Suzuki, H., Kainz, B., Matthews, P.M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Glocker, B., Rueckert, D., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J. Cardiovasc. Magn. Reson. 20 (1), 65.

Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S.E., Piechnik, S.K., Neubauer, S., Evangelou, E., Dehghan, A., O'Regan, D.P., Wilkins, M.R., Guo, Y., Matthews, P.M., Rueckert, D., 2020. A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Med. 26 (10), 1654–1662.

Bamberg, F., Kauczor, H.U., Weckbach, S., Schlett, C.L., Forsting, M., Ladd, S., Greiser, K.H., Weber, M.-A., Schulz-Menger, J., Niendorf, T., Pischon, T., Caspers, S., Amunts, K., Berger, K., Bülow, R., Hosten, N., Hegenscheid, K., Kröncke, T.J., Linseisen, J., Günther, M., Hirsch, J.G., Köhn, A., Hendel, T., Wichmann, H.-E., Schmidt, B., Jöckel, K.-H., Hoffmann, W., Kaaks, R., Reiser, M., Völzke, H., 2015. Whole-body MR imaging in the German National Cohort: Rationale, design, and technical background. Radiology 277 1, 206–220.

Bamberg, F., Schlett, C.L., Caspers, S., Ringhof, S., Günther, M., Hirsch, J.G., Rüdebusch, J., Miklánková, P., Bittner, N., Jockwitz, C., Forsting, M., Hosten, N., Kaaks, R., Kauczor, H.-U., Kroenke, T., Niendorf, T., Peters, A., Pischon, T., Stang, A., Berger, K., Völzke, H., 2024. Baseline MRI examination in the NAKO health study—Findings on feasibility, participation and dropout rates, comfort, and image quality. Dtsch. Arzteblatt Int. 121 (18), 587–593.

Basty, N., Liu, Y., Cule, M., Thomas, E.L., Bell, J.D., Whitcher, B., 2020. Automated measurement of pancreatic fat and iron concentration using multi-echo and T1-weighted MRI data. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, pp. 345–348.

Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M., 2018. The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9368–9377.

Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E., 2023. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proc. Natl. Acad. Sci. 120 (9), e2216399120.

Chen, P.-T., Wu, T., Wang, P., Chang, D., Liu, K.-L., Wu, M.-S., Roth, H.R., Lee, P.-C., Liao, W.-C., Wang, W., 2023. Pancreatic cancer detection on CT scans with deep learning: A nationwide population-based study. Radiology 306 (1), 172–182.

Elghazaly, H., McCracken, C., Szabo, L., Malcolmson, J., Manisty, C.H., Davies, A.H., Piechnik, S.K., Harvey, N.C., Neubauer, S., Mohiddin, S.A., 2023. Characterizing the hypertensive cardiovascular phenotype in the UK Biobank. Eur. Hear. Journal-Cardiovascular Imaging.

Ferdian, E., Suinesiaputra, A., Fung, K., Aung, N., Lukaschuk, E., Barutcu, A., Maclean, E., Paiva, J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Young, A.A., 2020. Fully automated myocardial strain estimation from cardiovascular MRI–tagged images using a deep learning framework in the UK Biobank. Radiol.: Cardiothorac. Imaging 2 (1).

Fischer, M., Walter, S.S., Hepp, T., Zimmer, M., Notohamiprodjo, M., Schick, F., Yang, B., 2021. Automated morphometric analysis of the hip joint on MRI from the German National Cohort study. Radiol.: Artif. Intell. 3 (5).

Galati, F., Zuluaga, M.A., 2021. Efficient model monitoring for quality control in cardiac image segmentation. In: Functional Imaging and Modeling of the Heart.

Geifman, Y., El-Yaniv, R., 2017. Selective classification for deep neural networks. Adv. Neural Inf. Process. Syst. 30.

Hann, E., Popescu, I.A., Zhang, Q., Gonzales, R.A., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K., 2021. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Med. Image Anal. 71, 102029.

Haueise, T., Schick, F., Stefan, N., Schlett, C.L., Weiss, J.B., Nattenmüller, J., Göbel-Guéniot, K., Norajitra, T., Nonnenmacher, T., Kauczor, H.-U., Maier-Hein, K.H., Niendorf, T., Pischon, T., Jöckel, K.-H., Umutlu, L., Peters, A., Rospleszcz, S., Kröncke, T., Hosten, N., Völzke, H., Krist, L., Willich, S.N., Bamberg, F., Machann, J., 2023. Analysis of volume and topography of adipose tissue in the trunk: Results of MRI of 11,141 participants in the German National Cohort. Sci. Adv. 9 (19).

Hepp, T., Blum, D., Armanious, K., Schölkopf, B., Stern, D., Yang, B., Gatidis, S., 2021. Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the German National Cohort MRI study. Comput. Med. Imaging Graph. 92, 101967.

Hepp, T., Fischer, M., Winkelmann, M.T., Baldenhofer, S., Kuestner, T., Nikolaou, K., Yang, B., Gatidis, S., 2020. Fully automated segmentation and shape analysis of the thoracic aorta in non–contrast-enhanced magnetic resonance images of the German National Cohort study. J. Thorac. Imaging 35 (6), 389–398.

Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H., Kalpathy-Cramer, J., 2020. An exploration of uncertainty information for segmentation quality assessment. In: Medical Imaging 2020: Image Processing, vol. 11313, pp. 381–390.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jamialahmadi, O., Tavaglione, F., Rawshani, A., Ljungman, C., Romeo, S., 2023. Fatty liver disease, heart rate and cardiac remodelling: Evidence from the UK Biobank. Liver Int. 43 (6), 1247–1255.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front. Neurosci. 14, 282.

Kalkhof, J., Mukhopadhyay, A., 2023. M3D-NCA: Robust 3D segmentation with built-in quality control. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, vol. 14222, pp. 169–178.

Karch, A., Vogelmeier, C., Welte, T., Bals, R., Kauczor, H.-U., Biederer, J., Heinrich, J., Schulz, H., Gläser, S., Holle, R., Watz, H., Korn, S., Adaskina, N., Biertz, F., Vogel, C., Vestbo, J., Wouters, E.F., Rabe, K.F., Söhler, S., Koch, A., Jörres, R.A., 2016. The German COPD cohort COSYCONET: Aims, methods and descriptive analysis of the study population at baseline. Respir. Med. 114, 27–37.

Kart, T., Fischer, M., Winzeck, S., Glocker, B., Bai, W., Bülow, R., Emmel, C., Friedrich, L., Kauczor, H.-U., Keil, T., Kröncke, T., Mayer, P., Niendorf, T., Peters, A., Pischon, T., Schaarschmidt, B.M., Schmidt, B., Schulze, M.B., Umutle, L., Völzke, H., Küstner, T., Bamberg, F., Schölkopf, B., Rueckert, D., Gatidis, S., 2022.

Automated imaging-based abdominal organ segmentation and quality control in 20,000 participants of the UK Biobank and German National Cohort studies. Sci. Rep. 12 (1), 18733.

Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L., 2012. Evaluating segmentation error without ground truth. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012, vol. 7510, Springer Berlin Heidelberg.

Konietzke, P., Weinheimer, O., Triphan, S.M., Nauck, S., Wuennemann, F., Konietzke, M., Jobst, B.J., Jörres, R.A., Vogelmeier, C.F., Heussel, C.P., et al., 2025. GOLD grade-specific characterization of COPD in the COSYCONET multi-center trial: comparison of semiquantitative MRI and quantitative CT. Eur. Radiol. 35 (7), 3842–3857.

Küstner, T., Hepp, T., Fischer, M., Schwartz, M., Fritsche, A., Häring, H.-U., Nikolaou, K., Bamberg, F., Yang, B., Schick, F., Gatidis, S., Machann, J., 2020. Fully automated and standardized segmentation of adipose tissue compartments via deep learning in 3D whole-body MRI of epidemiologic cohort studies. Radiol.: Artif. Intell. 2 (6).

Langner, T., 2021. Deep regression and segmentation for medical inference from large-scale magnetic resonance imaging.

Langner, T., Gustafsson, F.K., Avelin, B., Strand, R., Ahlström, H., Kullberg, J., 2021. Uncertainty-aware body composition analysis with deep regression ensembles on UK Biobank MRI. Comput. Med. Imaging Graph. 93, 101994.

Langner, T., Hedström, A., Mörwald, K., Weghuber, D., Forslund, A., Bergsten, P., Ahlström, H., Kullberg, J., 2019a. Fully convolutional networks for automated segmentation of abdominal adipose tissue depots in multicenter water–fat MRI. Magn. Reson. Med. 81 (4), 2736–2745.

Langner, T., Strand, R., Ahlström, H., Kullberg, J., 2020. Large-scale biometry with interpretable neural network regression on UK Biobank body MRI. Sci. Rep. 10 (1), 17752.

Langner, T., Wikström, J., Bjerner, T., Ahlström, H., Kullberg, J., 2019b. Identifying morphological indicators of aging with neural networks on large-scale whole-body MRI. IEEE Trans. Med. Imaging 39 (5), 1430–1437.

Lennartz, J., Schultz, T., 2025. Adversarial perturbations improve generalization of confidence prediction in medical image segmentation. In: Medical Imaging with Deep Learning.

Li, K., Yu, L., Heng, P.-A., 2022. Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references. Med. Image Anal. 78, 102426.

Linmans, J., Elfwing, S., van der Laak, J., Litjens, G.J.S., 2022. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med. Image Anal. 83.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Littlejohns, T.J., Holliday, J., Gibson, L.M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J.D., Boultwood, C., Collins, R., Conroy, M.C., 2020. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. 11 (1), 2624.

Liu, Y., Basty, N., Whitcher, B., Bell, J.D., Sorokin, E.P., van Bruggen, N., Thomas, E.L., Cule, M., 2021. Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. Elife 10, e65554.

Liu, F., Xia, Y., Yang, D., Yuille, A.L., Xu, D., 2019. An alarm system for segmentation algorithm based on shape model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10652–10661.

Mahmood, A., Simon, J., Cooper, J., Murphy, T., McCracken, C., Quiroz, J., Laranjo, L., Aung, N., Lee, A.M., Khanji, M.Y., 2023. Neuroticism personality traits are linked to adverse cardiovascular phenotypes in the UK Biobank. Eur. Hear. Journal-Cardiovascular Imaging 24 (11), 1460–1467.

Mariscal-Harana, J., Asher, C., Vergani, V., Rizvi, M., Keehn, L., Kim, R.J., Judd, R.M., Petersen, S.E., Razavi, R., King, A.P., 2023. An artificial intelligence tool for automated analysis of large-scale unstructured clinical cine cardiac magnetic resonance databases. Eur. Hear. Journal-Digital Health 4 (5), 370–383.

Mauger, C., Gilbert, K., Lee, A.M., Sanghvi, M.M., Aung, N., Fung, K., Carapella, V., Piechnik, S.K., Neubauer, S., Petersen, S.E., Suinesiaputra, A., Young, A.A., 2019. Right ventricular shape and function: cardiovascular magnetic resonance reference morphology and biventricular risk factor morphometrics in UK Biobank. J. Cardiovasc. Magn. Reson. 21 (1), 41.

Ng, M., Guo, F., Biswas, L., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G., 2022. Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study. IEEE Trans. Biomed. Eng..

Peters, A., Peters, A., Greiser, K.H., Göttlicher, S., Ahrens, W., Albrecht, M., Bamberg, F., Bärnighausen, T.W., Becher, H., Berger, K., Beule, A.G., Boeing, H., Bohn, B., Bohnert, K., Braun, B., Brenner, H., Bülow, R., Castell, S., Damms-Machado, A., Dörr, M., Ebert, N., Ecker, M., Emmel, C., Fischer, B., Franzke, C.-W., Gastell, S., Giani, G., Günther, M., Günther, K., Günther, K.P., Haerting, J., Haug, U., Heid, I.M., Heier, M., Heinemeyer, D., Hendel, T., Herbolsheimer, F., Hirsch, J.G., Hoffmann, W., Holleczek, B., Hölling, H., Hörlein, A., Jöckel, K.-H., Kaaks, R., Karch, A., Karrasch, S., Kartschmit, N., Kauczor, H.U., Keil, T., Kemmling, Y., Klee, B., Klüppelholz, B., Kluttig, A., Kofink, L.S., Köttgen, A., Kraft, D.N., Krause, G., Kretz, L., Krist, L., Kühnisch, J., Kuss, O., Legath, N., Lehnich, A.-T., Leitzmann, M.F., Lieb, W., Linseisen, J., Loeffler, M., Macdonald, A., Maier-Hein, K.H., Mangold, N., Meinke-Franze, C., Meisinger, C., Melzer, J., Mergarten, B., Michels, K.B., Mikolajczyk, R.T., Moebus, S., Mueller, U., Nauck, M., Niendorf, T., Nikolaou, K., Obi, N., Ostrzinski, S., Panreck, L., Pigeot, I., Pischon, T., Pschibul-Thamm, I., Rathmann, W., Reineke, A., Roloff, S., Rujescu, D., Rupf, S., Sander, O., Schikowski, T., Schipf, S., Schirmacher, P., Schlett, C.L., Schmidt, B., Schmidt, G., Schmidt, M., Schöne, G., Schulz, H., Schulze, M.B., Schweig, A., Sedlmeier, A.M., Selder, S., Six-Merker, J., Sowade, R.F., Stang, A., Stegle, O., Steindorf, K., Stübs, G., Swart, E., Teismann, H., Thiele, I., Thierry, S., Ueffing, M., Völzke, H., Waniek, S., Weber, A., Werner, N., Wichmann, H.-E., Willich, S.N., Wirkner, K., Wolf, K., Wolff, R., Zeeb, H., Zinkhan, M., Zschocke, J., 2022. Framework and baseline examination of the German National Cohort (NAKO). Eur. J. Epidemiol. 37, 1107–1124.

Pirruccello, J.P., Di Achille, P., Nauffal, V., Nekoui, M., Friedman, S.F., Klarqvist, M.D., Chaffin, M.D., Weng, L.-C., Cunningham, J.W., Khurshid, S., 2022. Genetic analysis of right heart structure and function in 40,000 people. Nature Genet. 54 (6), 792–803.

Qiu, P., Chakrabarty, S., Nguyen, P., Ghosh, S.S., Sotiras, A., 2023. QCResUNet: Joint subject-level and voxel-level prediction of segmentation quality. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, vol. 14223, pp. 173–182, Series Title: Lecture Notes in Computer Science.

Raisi-Estabragh, Z., Biasiolli, L., Cooper, J., Aung, N., Fung, K., Paiva, J.M., Sanghvi, M.M., Thomson, R.J., Curtis, E., Paccou, J., Rayner, J.J., Werys, K., Puchta, H., Thomas, K.E., Lee, A.M., Piechnik, S.K., Neubauer, S., Munroe, P.B., Cooper, C., Petersen, S.E., Harvey, N.C., 2021. Poor bone quality is associated with greater arterial stiffness: Insights from the UK Biobank. J. Bone Miner. Res. 36 (1), 90–99.

Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Kainz, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Rueckert, D., Glocker, B., 2018. Real-time prediction of segmentation quality. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, vol. 11073, pp. 578–585.

Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Matthews, P.M., Rueckert, D., Glocker, B., 2019. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. J. Cardiovasc. Magn. Reson. 21 (1), 18.

Robinson, R., Valindria, V.V., Bai, W., Suzuki, H., Matthews, P.M., Page, C., Rueckert, D., Glocker, B., 2017. Automatic quality control of cardiac MRI segmentation in large-scale population imaging. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017, vol. 10433, pp. 720–727.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., 2018. Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, vol. 11070, pp. 664–672.

Schübel, R., Graf, M.E., Nattenmüller, J., Nabers, D., Sookthai, D., Gruner, L.F., Johnson, T., Schlett, C.L., von Stackelberg, O., Kirsten, R., Habermann, N., Kratz, M., Kauczor, H.U., Ulrich, C.M., Kaaks, R., Kühn, T., 2016. The effects of intermittent calorie restriction on metabolic health: Rationale and study design of the HELENA trial. Contemp. Clin. Trials 51, 28–33.

Settles, B., 2009. Active learning literature survey. URL: https://api.semanticscholar.org/CorpusID:324600.

Thanaj, M., Basty, N., Cule, M., Sorokin, E.P., Whitcher, B., Srinivasan, R., Lennon, R., Bell, J.D., Thomas, E.L., 2023. Kidney shape statistical analysis: associations with disease and anthropometric factors. BMC Nephrol. 24 (1), 362.

Thanaj, M., Basty, N., Liu, Y., Cule, M., Sorokin, E.P., Louise Thomas, E., Bell, J.D., Whitcher, B., 2021. Mass univariate regression analysis for three-dimensional liver image-derived phenotypes. In: Medical Image Understanding and Analysis, vol. 12722, pp. 165–176.

Triay Bagur, A., Aljabar, P., Ridgway, G.R., Brady, M., Bulte, D.P., 2022. Pancreas MRI segmentation into head, body, and tail enables regional quantitative analysis of heterogeneous disease. J. Magn. Reson. Imaging 56 (4), 997–1008.

Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W., 2020. Deep generative model-based quality control for cardiac MRI segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, vol. 12264, pp. 88–97.

Whitcher, B., Thanaj, M., Cule, M., Liu, Y., Basty, N., Sorokin, E.P., Bell, J.D., Thomas, E.L., 2022. Precision MRI phenotyping enables detection of small changes in body composition for longitudinal cohorts. Sci. Rep. 12 (1), 3748.

Wolf, I., Vetter, M., Wegner, I., Nolden, M., Bottger, T., Hastenteufel, M., Schobinger, M., Kunert, T., Meinzer, H.-P., 2004. The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK. In: Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display, vol. 5367, pp. 16–27.

Wu, T., Estrada, S., Van Gils, R., Su, R., Jaddoe, V.W.V., Oei, E.H.G., Klein, S., 2023. Automated deep learning–based segmentation of abdominal adipose tissue on dixon MRI in adolescents: A prospective population-based study. Am. J. Roentgenol. 1–12.

Yazdan-Panah, A., Stankoff, B., Colliot, O., 2024. Automatic quality control of segmentation results using early epochs as data augmentation: application to choroid plexuses. In: Medical Imaging.

Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P.F., Maier-Hein, K., 2025. Comparative benchmarking of failure detection methods in medical image segmentation: unveiling the role of confidence aggregation. Med. Image Anal. 101, 103392.

Zheng, Q., Delingette, H., Fung, K., Petersen, S.E., Ayache, N., 2020. Pathological cluster identification by unsupervised analysis in 3,822 UK Biobank cardiac MRIs. Front. Cardiovasc. Med. 7, 539788.