

Pancreas MRI segmentation into head, body, and tail enables regional quantitative analysis of heterogeneous disease

Alexandre Triay Bagur MSc^{1,2*}, Paul Aljabar PhD², Gerard R Ridgway PhD², Prof. Michael Brady², and Prof. Daniel P Bulte¹

¹Department of Engineering Science, University of Oxford, Oxford, UK

²Perspectum Ltd, Oxford, UK

*Corresponding author: alexandre.triaybagur@eng.ox.ac.uk, University of Oxford, Old Road Campus Research Building, Headington, Oxford OX3 7DQ, +441865617675

Acknowledgements

The authors thank Joel Robinson for help with delineations and feedback on the annotation protocol. We thank Perspectum Ltd for funding and computational resources. We also thank the Engineering and Physical Sciences Research Council (EPSRC) for the doctoral studentship award. This research was conducted using the UK Biobank Resource under Application Number 9914.

Running Title

Pancreas head, body, tail segmentation

Keywords

Segmentation, groupwise registration, NAFPD, diabetes, MRI-PDFF, heterogeneity

Abstract

Pancreatic disease can be spatially inhomogeneous. For this reason, quantitative imaging studies of the pancreas have often targeted the 3 main anatomical pancreatic segments, head, body, and tail, traditionally using a balanced region of interest (ROI) strategy. Existing automated analysis methods have implemented whole-organ segmentation, which provides an overall quantification, but fails to address spatial heterogeneity in disease. A method to automatically refine a whole-organ segmentation of the pancreas into head, body, and tail subregions is presented for abdominal magnetic resonance imaging (MRI). The subsegmentation method is based on diffeomorphic registration to a group average template image, where the parts are manually annotated. For a new whole-pancreas segmentation, the aligned template's part labels are automatically propagated to the segmentation of interest. The method is validated retrospectively on the UK Biobank imaging substudy (scanned using a 2-point Dixon protocol at 1.5 tesla), using a nominally healthy cohort of 100 subjects for template creation, and 50 independent subjects for validation. Pancreas head, body, and tail were annotated by multiple experts on the validation cohort, which served as the benchmark for the automated method's performance. Good intra-rater (Dice overlap mean, Head: 0.982, Body: 0.940, Tail: 0.961, N=30) as well as inter-rater (Dice overlap mean, Head: 0.968, Body: 0.905, Tail: 0.943, N=150) agreement was observed. No significant difference (Wilcoxon rank sum test, DSC, Head: $p=0.4358$, Body: $p=0.0992$, Tail: $p=0.1080$) was observed between the manual annotations and the automated method's predictions. Results on regional pancreatic fat assessment are also presented, by intersecting the 3-D parts segmentation with one 2-D multi-echo gradient-echo slice, available from the same scanning session, that was used to compute MRI proton density fat fraction (MRI-PDFF). Initial application of the method on a type 2 diabetes cohort showed the utility of the method for assessing pancreatic disease heterogeneity.

1. Introduction

1.1. Heterogeneity of chronic pancreas disease

Pancreas pathology, such as fatty infiltration, diabetes, chronic pancreatitis, and pancreatic cancer, are rising rapidly with the increasing prevalence of obesity and the metabolic syndrome. Obesity leads to ectopic fat deposition in organs like the heart, liver, and pancreas. While non-alcoholic fatty liver disease (NAFLD) is a well-recognized disease entity, now affecting 1/4th of the worldwide population and 1/3rd of US adults [1], non-alcoholic fatty pancreas disease (NAFPD) was only coined relatively recently [2], [3] despite showing similar prevalence in a meta-analysis [4]. Analogously to NAFLD, NAFPD triggers inflammatory processes that, if left untreated, may lead to chronic pancreatitis and pancreatic cancer [5], [6]. NAFPD has also been linked to type 2 diabetes [7], [8]. Early detection of pancreatic disease is therefore important, however these are often “silent” conditions that only become symptomatic at a late stage, when they may already be untreatable. Incidental findings, where the target organ is near the pancreas, for instance in quantitative imaging of the liver, potentially offer a way to detect pancreas pathology in time.

Pancreatic disease processes, including fat infiltration, fibro-inflammation, and pancreatic cancer, are also spatially inhomogeneous [9], [10]. There is increasing interest in studying pancreatic disease and the implications of disease heterogeneity, aiming to describe regional differences and localize pancreatic lesions. Early work using computed tomography (CT) classified uneven pancreatic fat infiltration into multiple subtypes or patterns, depending on the affected regions [10]. Uneven distribution of islet cells, that are responsible for insulin secretion and blood sugar regulation, has been reported using histology [11]. Fibrosis was more commonly found in the ventral pancreas than in the dorsal pancreas in patients with ampullary carcinoma [12]. The frequency of pancreatic cancer also differs regionally, with 60-70% occurrence in the head of the pancreas, and the symptoms vary by the location [13], [14]. From the imaging modalities commonly used for pancreatic assessment, including histology, endoscopic ultrasound, contrast-enhanced CT, and magnetic resonance imaging (MRI), only MRI can provide non-ionizing, non-invasive quantitative information of pancreas state, while providing full coverage and measures of spatial heterogeneity. Quantitative MRI biomarkers such as proton density fat fraction (PDFF) and T₁ have shown potential in detecting pancreas steatosis and early-stage chronic pancreatitis, respectively [15], [16]; PDFF has been used for longitudinally monitoring total pancreatic fat deposition in a diabetes remission trial [17].

Apparent diffusion coefficient (ADC) from diffusion-weighted imaging has shown potential at grading malignancy of a certain pancreatic neoplasm type [18]. While some studies using MRI reported clinically important quantitative differences between pancreas subsegments [19], [20], other studies did not find such differences [21].

The pancreas is anatomically divided into three segments: head, body, and tail. The pancreas head sits within a C-shape structure formed by the duodenum and joins with the pancreas body via the pancreas neck, a narrowing or ‘isthmus’ that bends around the superior mesenteric vessels. The pancreas neck is typically approximately 2 cm long and is commonly included as part of the head. The pancreas body spans from the left border of the superior mesenteric vein to the left border of the aorta, where it is joined to the tail. It is generally considered that the body-tail boundary is at the midpoint lengthwise of the two segments [22]. Other pancreas subsegment classification systems have been proposed for the purposes of surgical resection, based upon embryological foundations [22], [23]. Most studies of pancreas pathology using MRI have analyzed the images using regions of interest (ROIs), particularly a standard 3-ROI placement strategy targeting pancreatic head, body, and tail [20], [21], [24], [25], though some have placed an extra ROI in the pancreatic neck [19]. While ROIs have the advantage of avoiding artefactual regions, their choice of placement inevitably adds inter-observer variation that may obscure clinically important differences between pancreatic segments.

1.2. Pancreas subsegmentation

Pancreas segmentation, that aims to delineate the whole organ in 2-D or 3-D scans, has been proposed as an alternative analysis method to the 3-ROI placement strategy, which may improve observer-dependent bias and provide more advanced metrics for spatial assessment of chronic disease. However, such is the variability in size and shape of the pancreas that it is often considered too tedious to manually delineate in practice. The methods for pancreas segmentation proposed to date require widely differing amounts of user intervention. Manual segmentation is too costly and generally infeasible, especially in large databases such as the UK Biobank [26]. Metrics derived from pancreas segmentations are clinically important, for instance total pancreatic volume or the irregularity of the pancreas contour in the context of diabetes [27], [28]. Pancreas segmentations may also be used for subsequent characterization of the pancreas in functional or structural quantitative imaging data acquired separately during the same imaging session.

Automated pancreas segmentation methods that have been proposed to date have been based on traditional multi-atlas methodology or, more recently, convolutional neural networks [29]–[31]. The latter especially have reported remarkable accuracy. However, while these may provide whole-organ measurements, they do not characterize disease regionally by pancreas subsegments. One automated method for pancreas subsegmentation was reported based on k-means clustering [32], that was applied to pancreas motion analysis under radiation therapy. However, this method is dependent on initial seed points and multiple images from multiple breathing phases, and was not validated for accuracy. For these reasons, the validation of a robust, automated approach for pancreas subsegmentation is desirable, with potential to bridge the gap between currently available technology and standard clinical assessment.

Starting from a segmentation of a whole-organ, landmark-based approaches have been proposed for subsegmentation into the organ’s constituent parts, for example the Couinaud segments in the case of the liver [33]. In one approach, “landmarks” have been used to define planes of separation between the liver segments. However, landmark localization is relatively sensitive to noise and overall image quality. Other methods have addressed organ subsegmentation as a single task, in which segmentation models create a multi-label prediction, each label corresponding to an individual subsegment. For example, atlas-based segmentation uses image registration to propagate labels from a probabilistic template (constructed offline) to a target dataset. Multi-atlas segmentation (MAS) or deep learning (DL) segmentation methods may also be used, however, they typically need annotated parts individually on training subjects, and generally require large amounts of data. Some DL methods have drawn inspiration from traditional atlas-based methodology [34].

1.3. Purpose

In this work, a fully automated method based on groupwise registration is presented to subsegment the pancreas into its anatomical parts. Automatic delineations are assessed in comparison with those of human experts. Intra- and inter-rater variation of human experts is determined, that sets a benchmark for assessment of the automated method. We show that we can automatically segment pancreas parts in MRI, starting from a whole-pancreas segmentation, further enabling work studying disease heterogeneity. We validate the method retrospectively on a subset from the UK Biobank.

The purpose of this study is three-fold:

1. Propose a robust annotation protocol for delineating pancreas subsegments.
2. Introduce and validate an automated method for pancreas subsegmentation.
3. Show initial application of the method in regional assessment of pancreatic disease.

2. Materials and Methods

In Section 2.1, the data that were used for template creation are described, together with pre-processing of the training and validation data. The groupwise registration-based parts segmentation method is described in Section 2.2. In Section 2.3, we describe the validation experiment using expert annotations as reference. Section 2.4 shows the application of the method to a type 2 diabetes cohort of UK Biobank.

As benchmark, the method based on k-means proposed by Fontana et al. [32] was implemented for a single image (single ‘breathing phase’), choosing the initial cluster centroids using the k-means++ algorithm, and is referred to as the *k-means method* throughout this paper.

2.1. MRI data

MRI data from the UK Biobank imaging substudy was used. UK Biobank received ethical approval from the North West Multi-centre Research Ethics Committee (MREC) and written informed consent was obtained for all subjects. 100 subjects were used for template creation, 44 females and 56 males. All were nominally healthy subjects aged 50 to 70 with a mean age of 55 years for females and 57 years for males. The mean Body Mass Index (BMI) was 25.5 kg/m² for females and 27.1 kg/m² for males. An additional 50 subjects were used for validation, 21 males and 29 females, with a mean age of 53 and 57 years and a BMI of 25.9 and 26.5 kg/m², respectively.

All subjects had been scanned with a 1.5 tesla Siemens Aera (Siemens Healthineers, Erlangen, Germany) using a 2-point Dixon protocol covering neck to knee, acquired using 6 overlapping slabs and uploaded as Data-Field 20201. Only datasets from the first imaging session of UK Biobank (Instance 2) were used. Slabs were stitched together, and the resulting neck-to-knee volume was cropped to the abdominal region, resulting in a subvolume that generally included

slabs 2, 3 and 4 (more details are available in [35]). Slabs 2, 3 and 4 each had voxel size 2.23 x 2.23 x 4.5 mm³ and matrix size 224 x 174 x 44.

The whole pancreas was delineated manually on all 150 training and validation datasets. Figure 1 shows 3-D renderings of the whole-pancreas segmentations for all subjects in both the template creation dataset and the validation dataset. The volumes and the corresponding whole-organ segmentations were resampled to 2 mm isotropic resolution. We also minimally co-registered the subjects by translating them to align their centroids. The centroid of subject 1 was used as a reference. The pre-alignment provided a better starting point for the nonlinear registration algorithm, both for template creation and method inference. Currently, the software is only compatible with isotropic images, identical image size and in approximate alignment with each other.

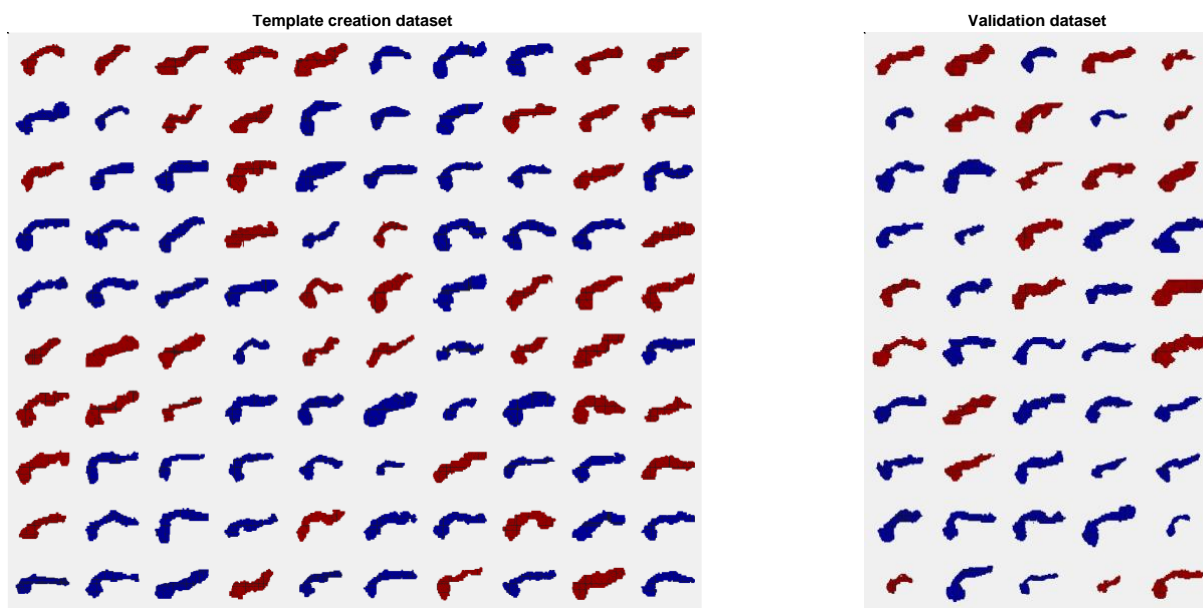


Figure 1. Whole-pancreas segmentations from the template construction (‘training’) dataset and the validation dataset, sorted by subjects’ age (females in red, males in blue).

2.2. Method description

An overview of the groupwise registration method is shown in Figure 2. The method takes a whole-pancreas segmentation as input, either delineated manually or with an automated approach. First, an average pancreas template is constructed offline using groupwise registration from the N=100 template creation dataset of whole-pancreas segmentations. Then, the pancreas parts (head, body, tail) are manually annotated on the constructed template,

resulting in a pancreas parts template. Method inference (parts segmentation) is performed by registration of the pancreas parts template to a new target whole-pancreas segmentation. Then, the registered parts template labels are propagated to the target whole-pancreas segmentation, obtaining a pancreas parts segmentation for that subject. Offline parts template construction as well as parts segmentation inference steps are detailed in the following paragraphs.

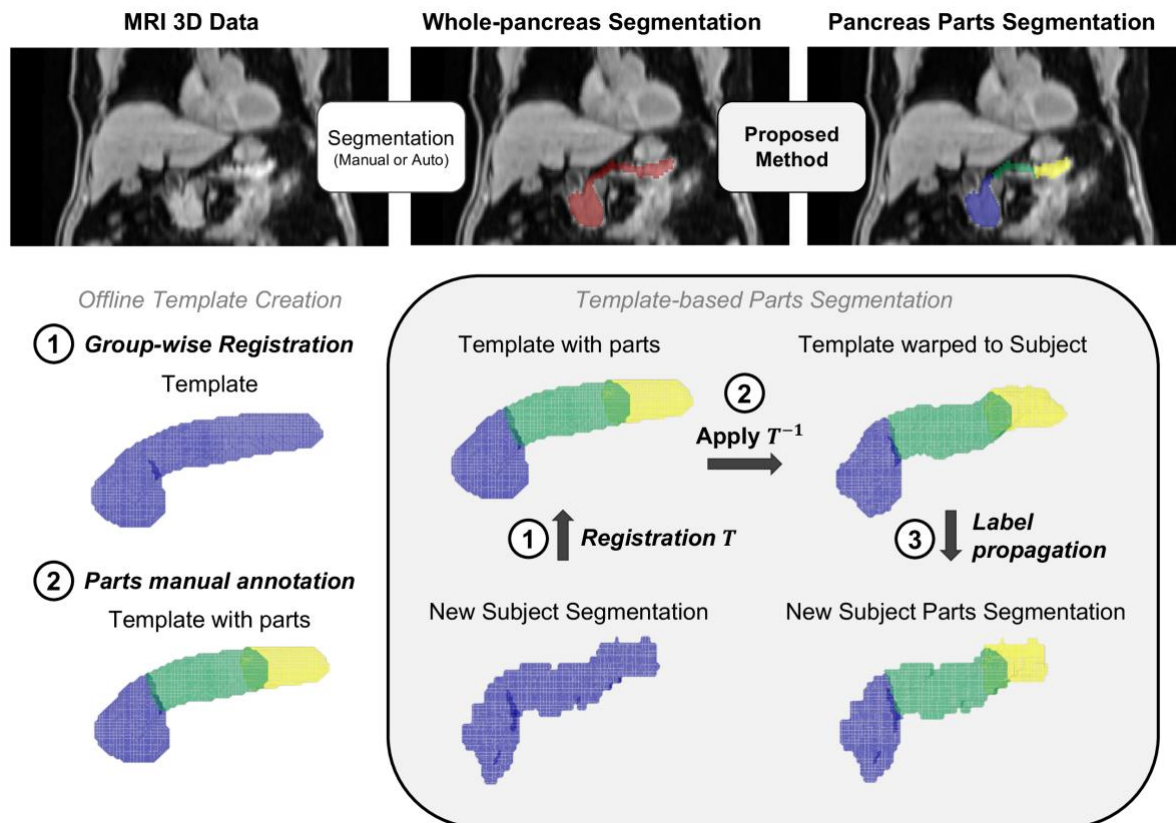


Figure 2. Method description. (top) Overall pipeline for whole-pancreas segmentation and parts segmentation. (bottom-left) (1) Offline groupwise registration of the whole-pancreas segmentation generated a population average ('template'), on which (2) the parts were manually annotated ('parts template', head: blue, body: green, tail: yellow). (bottom-right) For a new subject, the method (1) computes a registration transformation from the subject's segmentation to the template, (2) applies the inverse transformation on the parts template, and (3) propagates the warped parts template labels to the segmentation.

The backbone for template construction is the Large Deformation Diffeomorphic Metric Mapping (LDDMM) via Geodesic Shooting algorithm developed by [36] and available under

the “Shoot” toolbox of the SPM12 software¹. The toolbox uses diffeomorphic transformations to co-register all the template construction segmentations iteratively into a population average, i.e., the ‘template’ image. MATLAB R2021a (The MathWorks, Inc) and the batch processing capability of SPM12 were used to run template creation. A probabilistic template (0 to 1) was obtained from this step after 4 iterations, that was binarized by thresholding at 0.5.

Pancreas head, body, and tail were annotated on the template image. Note this template-based approach enables annotation of parts on the constructed template, instead of annotating each of the training subjects individually, thus requiring a single annotation step. The initial assumption was that this approach would not be significantly different from annotating each ‘training’ subject individually. One additional advantage of this annotation strategy is that some salient features appear on the template after groupwise registration, which correspond to the landmarks defining the pancreas subsegments. These landmarks may otherwise be difficult to identify in individual cases, and correct landmark identification is highly dependent on image quality. Annotation was performed by defining one boundary plane between head and body and another boundary plane between body and tail.

Given a whole-pancreas segmentation for a new subject, which can be either manually delineated or computed automatically, the method first computes a registration transformation from the subject’s whole-pancreas segmentation to the template (again initialized by aligning the centroid). The method then applies the inverse of that registration transformation onto the parts template. Finally, it propagates the labels of the warped parts template towards the whole-pancreas segmentation, obtaining a parts segmentation for that new subject.

2.3. Validation

A separate ‘validation’ dataset of N=50 subjects was gathered from UK Biobank with the same demographics as the template creation dataset. Automated pancreas subsegmentation was performed using the k-means method and the described groupwise registration method. Each subject’s whole-pancreas segmentation was also manually annotated into parts for reference, following a dedicated annotation protocol based on the 3-D ‘scalpel’ tool of ITK-SNAP² [37]. The protocol instructed the drawing of 2 separation planes, one plane at the head-body

¹ <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

² <http://www.itksnap.org/>

boundary and one plane at the body-tail boundary, both as perpendicular to the pancreas centerline as possible. The protocol was distributed to 4 separate raters. One of the raters, who we refer to as *naïve observer*, had no prior experience with pancreas anatomy or pancreas imaging, so that the robustness of the annotation protocol to rater experience could be estimated. The three other raters, all of whom had expertise in medical imaging and familiarity with pancreas anatomy, were deemed “expert”. 10 of the 50 subjects were included at random twice in the dataset for the purpose of assessing intra-observer variability (referred to as annotation *a* and annotation *b*). This yielded a total of 60 annotations per rater. Inter-observer variability was also assessed by comparing annotations over multiple raters. The inter-observer performance may be used as a comparative benchmark for the automatic results.

Direct validation of automated pancreas subsegmentation was performed using generally accepted segmentation performance metrics, namely Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (95%HD), as well as the reported volume of each part. Intra-observer agreement, inter-observer agreement and ‘manual vs automated’ agreement were evaluated using Bland-Altman analysis [38] and Wilcoxon signed rank statistical testing. For the 10 subjects used in intra-rater variation assessment, a total of 30 datapoints (10 subjects x 3 expert raters) were generated and combined for the comparison: R1a vs R1b, R2a vs R2b, R3a vs R3b. For inter-observer variation assessment, 3 comparisons among raters were combined, R1 vs R2, R1 vs R3, R2 vs R3, with 50 parts segmentations in each comparison, yielding 150 datapoints. For manual vs automated (*Auto*), the following comparisons were performed and combined for each automated method separately: R1 vs Auto, R2 vs Auto, R3 vs Auto, each with 50 parts segmentations, that resulted in 150 datapoints. The volume of individual parts was also compared for its potential clinical relevance as a biomarker.

Indirect validation was also performed through quantification of pancreatic fat by parts, using the multi-echo gradient-recalled echo (GRE) 2-D single-slice data obtained from a separate breath-hold scan, uploaded as Data-Field 20260. The GRE scan has 2.5 x 2.5 x 6 mm³ resolution, 160 x 160 matrix size, 10 echoes, $TE_1 = \Delta TE = 2.38$ ms, $TE_{10} = 23.8$ ms, $TR = 27$ ms, and 20° flip angle. The median values from head, body and tail were reported after reslicing the parts segmentation onto the reconstructed proton density fat fraction (PDFF) map. A confounder-corrected magnitude-based chemical-shift encoding method [39] was used to reconstruct PDFF maps from the raw 10-echo GRE data. A spectral model from liver fat was

used [40]. The 3-D parts segmentation volume was intersected with the 2-D PDFF map using the DICOM Reference Coordinate System information, as illustrated in Figure 3.

Quantification differences were reported between the automated parts segmentation and the manual parts segmentations for each subject. Agreement in quantification was evaluated using Bland-Altman analysis. Segment masks with area of ≤ 30 pixels were excluded from the comparisons after a quality control (QC) step. The median PDFF of the segment masks was reported after excluding pixels with values exceeding 50%, followed by morphological opening with a disk structuring element of 3 pixels in diameter. The 50% PDFF threshold aimed to exclude non-parenchymal pancreatic tissue, for example surrounding visceral adipose tissue that could have been introduced due to slight subject motion between breath-holds.

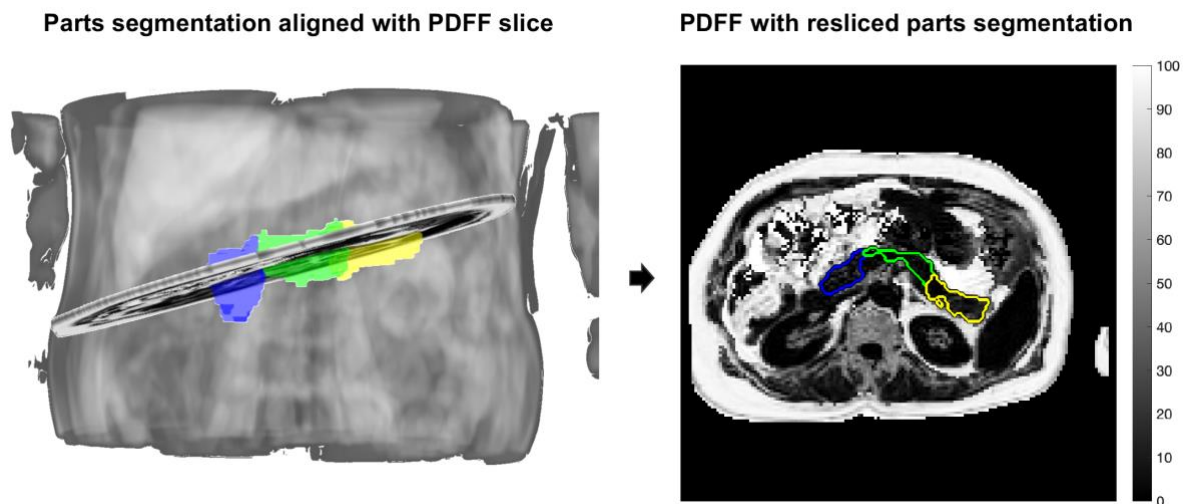


Figure 3. Pancreas segmentation (head: blue, body: green, tail: yellow) enables quantification of pancreas imaging biomarkers by parts, for example proton density fat fraction (PDFF), by intersection of the segmentation with the quantitative scan. PDFF, proton density fat fraction

2.4. Pancreatic fat quantification by parts in type 2 diabetes

As an initial exploration of fat heterogeneity in diseased subjects, a separate dataset of UK Biobank subjects was developed, comprising 390 triples of (1) self-reported type 2 diabetes mellitus (T2DM) subjects, (2) gender-, age- and BMI-matched non-diabetic subjects, and (3) gender- and age-matched non-diabetic subjects with chosen BMI of $< 25 \text{ kg/m}^2$. These groups of subjects will be referred to as: *T2DM*, *BMI-matched non-diabetics*, and *Matched low BMI non-diabetics* throughout this work. Age was matched to within 5 years, and BMI was matched within 1 point in all cases. A total of $390 \times 3 = 1,170$ subjects were collected.

Since manual whole-pancreas delineations were not available for these subjects, automated whole-pancreas segmentations obtained previously in [35] were used, predicted using the Attention U-Net model based on [31]. Some subjects had empty whole-pancreas segmentations that were excluded from the analysis. These were caused by failures in the model prediction, particularly in the presence of image artefacts. The groupwise registration-based automated parts segmentation method was run on the remaining whole-pancreas segmentations. The reslicing plus quality control approach explained in the previous section was run in order to measure median fat accumulation in the pancreatic head, body and tail. Pancreatic fat quantification by parts was compared between the 3 subject groups.

3. Results

3.1. Direct validation

Figure 4 illustrates manual and automated parts segmentations for the first 10 subjects in the validation set. 3-D renderings of manual parts segmentations from all 4 raters, as well as automated parts segmentations from the k-means method and the groupwise registration method are shown. The manual parts segmentations are similar across raters, including the naïve rater, with no major outliers. The manual parts segmentations look similar to the automated groupwise registration results, with no major outlier subjects, whereas the k-means method appears to overestimate the pancreatic head. These subjective judgements are quantified and confirmed below.

The robustness of the annotation protocol was tested against rater experience by comparing the segmentation performance in terms of DSC overlap of the naïve observer Rater (R) 4 vs themselves and vs the expert raters. The intra-rater agreement of a given expert observer R1 (R1a vs R1b) was not significantly higher than the intra-rater agreement of the naïve observer R4 (R4a vs R4b) (Wilcoxon signed rank test, right-tailed, Head: $p=0.3848$, Body: $p=0.2158$, Tail: $p=0.3125$). Similarly, the inter-observer agreement between 2 given expert observers, R1 vs R2, was not significantly higher than the inter-observer agreement between a given expert observer R1 vs the naïve observer R4 (Wilcoxon signed rank test, right-tailed, Head: $p=0.9827$, Body: $p=0.9486$, Tail: $p=0.9352$).

Table 1 reports intra-rater variation and inter-rater variation combined across the 3 expert raters and reported separately by pancreatic head, body and tail. Excellent intra-observer agreement (Dice overlap, Head: 0.982, Body: 0.940, Tail: 0.961, N=30) as well as inter-observer agreement (Dice overlap, Head: 0.968, Body: 0.905, Tail: 0.943, N=150) were observed in terms of segmentation performance. Intra-observer agreement was higher than inter-observer agreement.

Table 2 reports ‘manual vs automated’ differences combined across the 3 expert raters for both the k-means method and the groupwise registration method. For the k-means method, significant differences were found between the combined inter-observer agreement and the combined ‘manual vs automated’ agreement, using DSC (Wilcoxon rank sum test, DSC, Head: $p < 0.001$, Body: $p < 0.001$, Tail: $p = 0.3965$). For the groupwise registration method, no significant differences were found between the combined inter-observer agreement and the combined ‘manual vs automated’ agreement, using DSC (Wilcoxon rank sum test, DSC, Head: $p = 0.4358$, Body: $p = 0.0992$, Tail: $p = 0.1080$). A statistically significant difference was found between the ‘manual vs auto’ agreement of the k-means method and the ‘manual vs auto’ agreement of the groupwise registration method for the head (Wilcoxon signed rank test, DSC, $p < 0.001$) and body ($p < 0.001$) segments, but not for the tail ($p = 0.6237$).

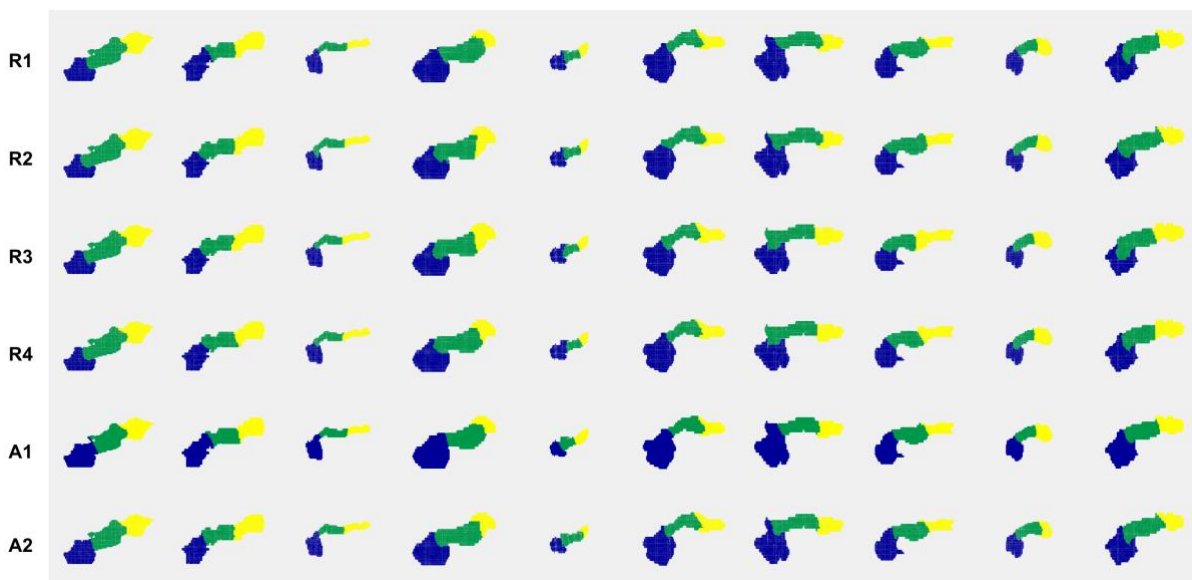


Figure 4. Qualitative evaluation of parts segmentations from Rater 1, Rater 2, Rater 3, Rater 4, Automated k-means method, Automated groupwise registration method, respectively (head: blue, body: green, tail: yellow). The first 10 subjects of the validation set are shown.

<i>Experts combined</i>	DSC	95%HD (mm)	ΔVolume (mL)
Intra-rater variation (n=10, N=30)			
head	0.982 \pm 0.014	0.609 \pm 1.342	0.19 [-1.83, 2.21]
body	0.940 \pm 0.033	2.934 \pm 2.420	0.08 [-3.95, 4.12]
tail	0.961 \pm 0.029	2.232 \pm 2.509	-0.28 [-4.25, 3.70]
Inter-rater variation (n=50, N=150)			
head	0.968 \pm 0.022	2.171 \pm 2.743	0.88 [-2.42, 4.19]
body	0.905 \pm 0.050	5.947 \pm 4.033	0.06 [-5.93, 6.05]
tail	0.943 \pm 0.042	4.026 \pm 4.121	-0.95 [-5.63, 3.73]

Table 1. Direct validation metrics, expert raters combined. Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (95%HD in mm) are reported as mean \pm standard deviation. Part volumes differences are reported in mL as bias [lower LoA, upper LoA]. Intra-observer agreement and inter-observer agreement are reported. LoA, limits of agreement

<i>Experts combined</i>	K-means method			Groupwise registration method		
	DSC	95%HD (mm)	ΔVolume (mL)	DSC	95%HD (mm)	ΔVolume (mL)
Manual vs Auto (n=50, N=150)						
head	0.942 \pm 0.045	6.196 \pm 5.096	-1.98 [-7.75, 3.79]	0.965 \pm 0.026	2.798 \pm 3.755	-0.63 [-4.46, 3.20]
body	0.855 \pm 0.120	8.899 \pm 5.855	0.59 [-4.50, 5.69]	0.893 \pm 0.058	6.397 \pm 3.890	0.08 [-5.69, 5.85]
tail	0.934 \pm 0.054	4.870 \pm 5.165	1.39 [-4.22, 7.00]	0.934 \pm 0.048	4.680 \pm 4.130	0.55 [-4.81, 5.91]

Table 2. Direct validation metrics, expert raters combined. Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (95%HD in mm) are reported as mean \pm standard deviation. Part volumes differences are reported in mL as bias [lower LoA, upper LoA]. Manual vs Automated agreement is reported for each of the existing automated methods. LoA, limits of agreement

3.2. Indirect validation

N=38 subjects had available the GRE sequence that enabled PDFF measurement out of the 50 subjects in the validation set. Note that, since the pancreatic PDFF scan is single-slice, the pancreatic head will not always present in the image due to variable slice positioning. Similarly, when slice position is too low, the pancreatic tail will not be visible. After processing and quality control, a total of 14 subjects with visible pancreatic head, 34 with visible body, and 29 with visible tail remained for quantification.

Figure 5 shows the Bland-Altman agreement in quantification for the inter-observer comparisons by pancreatic segments. Excellent agreement was observed between observers, for each part individually (head: bias=0.018, LoA=[-0.5, 0.5]; body: bias=-0.062, LoA=[-1.1, 1.0]; tail: bias=-0.019, LoA=[-1.6, 1.5]).

Figure 6 shows differences in parts quantification between manual parts segmentations and the quantification derived from the automated parts segmentations, reported separately by parts, for both the k-means method and the groupwise registration method. For the groupwise registration method, the agreement between manual and automated segmentation quantification was excellent and comparable to the inter-observer comparisons. For the k-means method, differences were observed between the ‘manual vs auto’ agreement and the inter-observer agreement, especially in the head segment (bias=-0.409, LoA=[-2.5, 1.7]).

Figure 7 shows an example of a subject’s PDFF map with the resliced parts segmentations from Rater 1, Rater 2, Rater 3, the groupwise registration method and the k-means method. The k-means method appeared to overestimate the head segment, which may have caused the quantitative differences observed in the ‘manual vs auto’ comparison.

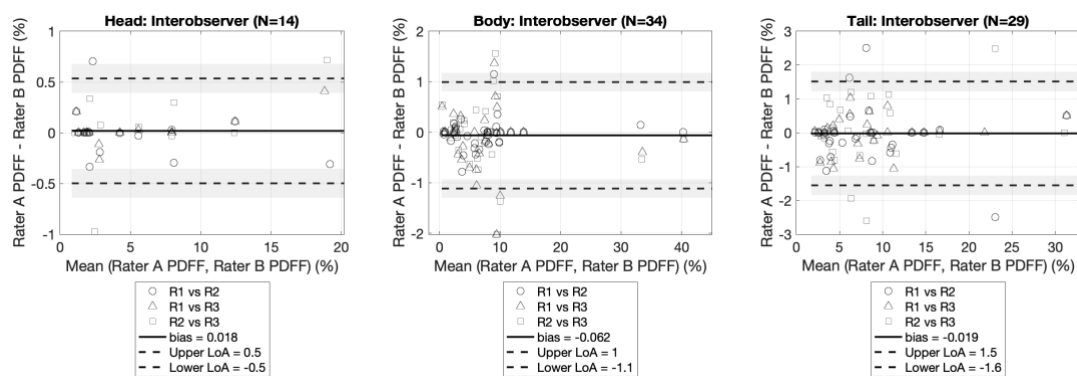


Figure 5. Inter-observer variation of derived PDFF quantification from the manual experts’ annotations.

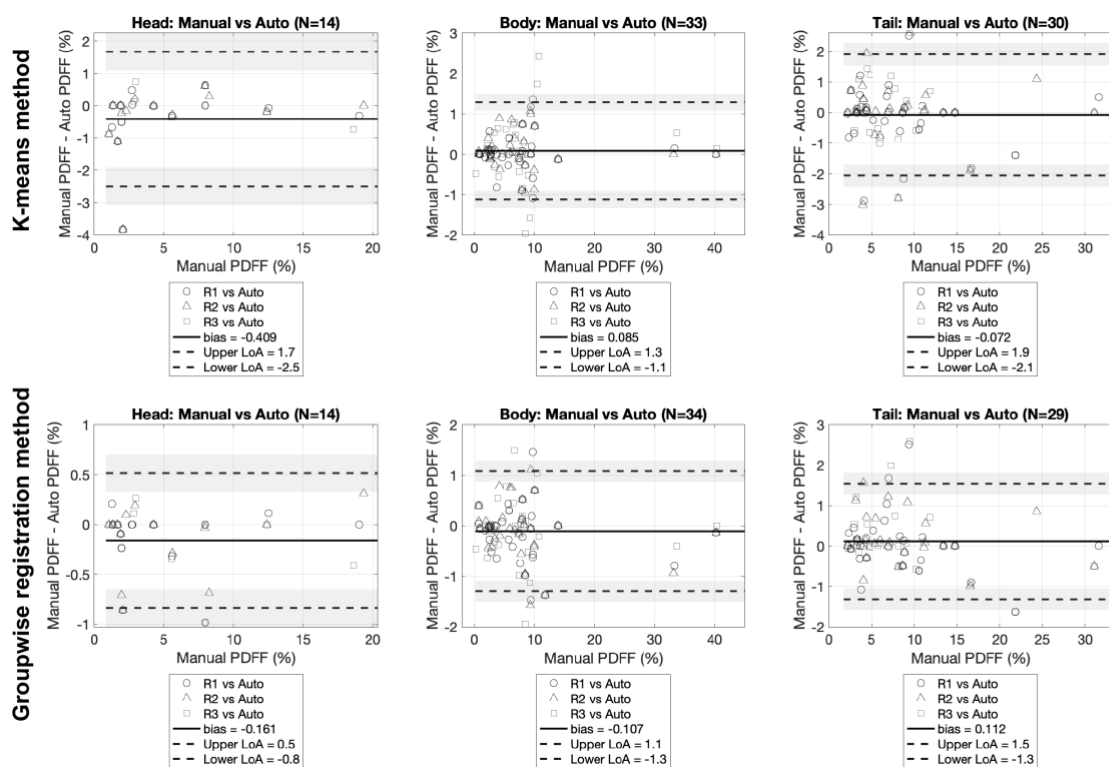


Figure 6. Manual experts' annotations vs automated subsegmentations-derived PDFF quantification: differences by parts. 'Manual vs auto' comparisons are presented for both the k-means method and the groupwise registration method.

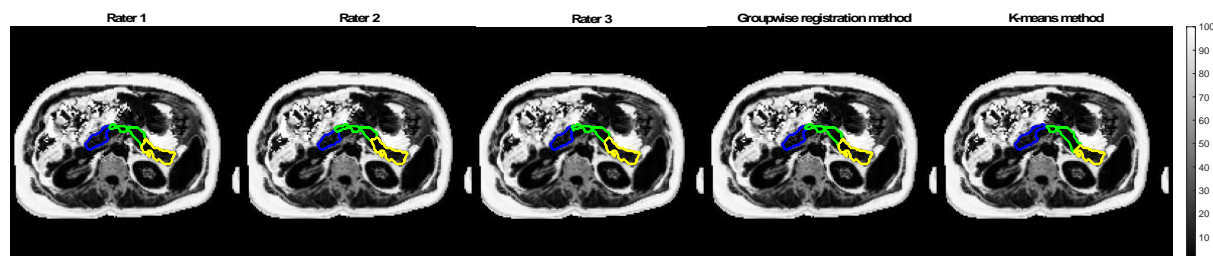


Figure 7. Example of regional pancreatic fat quantification for one subject in the validation set. Parts segmentation is shown on the PDFF image for Rater 1, Rater 2, Rater 3, Groupwise registration method, and K-means method, respectively (head: blue, body: green, tail: yellow). PDFF, proton density fat fraction

3.3. Pancreatic fat quantification by parts in type 2 diabetes

Figure 8 shows pancreatic fat quantification by pancreatic head, body, and tail for the 3 groups, (1) type 2 diabetics, (2) non-diabetics with matched age, gender and BMI, and (3) non-diabetics with matched age and gender and chosen low BMI. Total pancreatic fat is also included, which was obtained after combining all the part labels into a single 'whole' label.

The Mann-Whitney U-test was used to compare quantification by parts across groups of subjects. The difference in PDFF of parts between T2DM and matched non-diabetic subjects with low BMI was significant when comparing all whole-pancreas PDFF, head PDFF, body PDFF, and tail PDFF ($p < 0.001$ for all). The difference in PDFF of parts between T2DM and BMI-matched non-diabetics was only significant when comparing body PDFF ($p < 0.05$), but not significant when comparing whole-pancreas PDFF ($p = 0.067$), head PDFF ($p = 0.943$) or tail PDFF ($p = 0.623$).

Wilcoxon signed rank tests were used to evaluate differences in parts quantification between parts within each subject group. No significant difference between parts within groups was found, except for the BMI-matched cohort and the matched cohort with low BMI, where the head PDFF vs tail PDFF differences were significant ($p < 0.05$ for both).

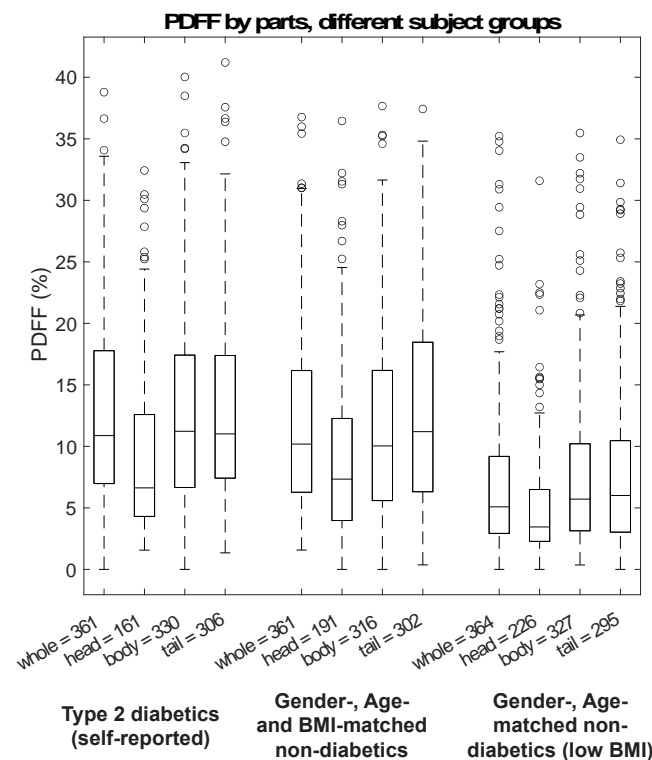


Figure 8. Pancreatic fat quantification by parts in groups from UK Biobank: (1) Type 2 diabetics (self-reported), (2) non-diabetics matched by gender, age, and BMI, (3) non-diabetics matched by gender and age and with low BMI. The number of segments available is displayed next to the part. BMI, body mass index

4. Discussion

This work presented and validated a fully automated method based on groupwise registration to subsegment the pancreas into its main anatomical parts: head, body, and tail. The method is based on a single population average or ‘template’ image and a single annotation stage on the template, which yields a parts template that may be used for pancreas subsegmentation in new subjects. The method was validated against manual annotations from expert observers in subjects from the UK Biobank imaging substudy and was compared to previously proposed unvalidated methodology based on k-means clustering [32]. Validation metrics included segmentation performance metrics as well as more clinically meaningful metrics like volume of parts and fat quantification by parts, which was obtained by intersecting the parts segmentations with proton density fat fraction (PDFF) maps. Then, as initial exploration of the clinical value of parts segmentation, the method was applied to a separate UK Biobank cohort including type 2 diabetics (self-reported) as well as gender-, age- and BMI-matched non-diabetic individuals, where spatial distribution of pancreatic PDFF was evaluated.

Note that automated whole-pancreas segmentation could have been used to generate both the template creation dataset and the validation dataset. However, by performing manual segmentations, any error introduced by whole-pancreas segmentation upstream of our method was minimized. This modular approach in which segmentation of the whole pancreas and the constituent parts are treated separately expedites validation of the subsegmentation method and allows for the introduction of improved whole-pancreas segmentation methods when they become available. In the final experiment, that showed the potential of parts segmentation, automated segmentations were used.

Excellent intra-rater as well as inter-rater agreement was observed among all raters for the proposed head, body, tail annotation protocol. This was true both for the 3 expert raters and the 1 ‘naïve’ rater, suggesting that the annotation protocol is robust, repeatable, and can be deployed by a range of raters.

Most literature quantifies imaging biomarkers by head, body, tail [20], [21], [24], as in the work presented here, though some researchers have considered the pancreatic neck separately in the quantification [19]. In early stages of developing the annotation protocol, four, rather than three, pancreas subsegments were considered by subdividing the head further into head

and neck. This resulted in significant variation between raters. In any case, considering the image resolution of the PDFF map in UK Biobank, the pancreatic neck area would be comprised of few pixels, diminishing the reliability of neck PDFF quantification. Other acquisitions and applications may be more suitable for separate neck quantification, which we will revisit in future work. Other pancreas subsegmentation systems, for instance those incorporating embryological basis [22], [23], should also be considered in the future, for they may provide complementary regional assessment of the pancreas.

Excellent agreement was observed between the manual annotations and the automated groupwise registration method's predictions, in terms of segmentation performance, subsegment volumes, and derived PDFF quantification. The agreement between expert raters and the automatic method suggests that the latter can be used in large datasets such as the UK Biobank. For this reason, the automated groupwise registration method was used in the subsequent experiment, which characterized regional quantification of fat in type 2 diabetics. One limitation of this is that template construction was performed using UK Biobank data comprising nominally healthy volunteers aged 50 to 70 with no self-reported diabetes of any type, though we plan to expand this cohort in future versions of the method. Applying the method to the type 2 diabetes cohort did not seem to impair method performance, based on a qualitative assessment (results not shown), though this needs careful evaluation.

As an initial exploration of the clinical application of our parts segmentation, we considered three matched groups: self-reported type 2 diabetes mellitus (T2DM) subjects, BMI-matched non-diabetics, and age- and gender-matched non-diabetics with low BMI. The significantly higher whole-pancreas PDFF in diabetics compared to non-diabetics has been reported previously [25]. However, we have shown that PDFF in the pancreatic body is significantly different between T2DM and BMI-matched non-diabetics, demonstrating the importance of parts segmentation beyond whole-pancreas measurements, which may obscure subtle but clinically important differences. One other study showed PDFF in the pancreatic tail to be most predictive for T2DM development within 4 years [20]. Our finding needs to be examined in more detail in future validation, for example using dedicated T2DM cohorts with longitudinal follow-up. The shown evidence of significant differences in pancreatic fat content between the pancreatic parts emphasizes the importance of segmentation-based approaches over ROI protocols, which should at least be 'balanced' when used, meaning they should target all pancreatic segments, for instance using multiple slices at different positions.

One advantage of groupwise registration methods is that they may be used for subsequent statistical analysis of biological variation across the population. Also, they often generalize robustly to various scan settings, compared to for instance deep learning methods, provided that the image resolution is in the same order. One criticism of templates is that they might average out differences between subjects. An approach that considers multiple templates based on major components of variation may be useful, for example clinical metadata information or imaging-based and radiomics features [29]. However, this increases the number of templates that need separate annotation. Evidently, in the extrema of this approach sit multi-atlas segmentation (MAS) methods, for which individual subjects in the training set need manual annotation of parts. Our approach seemed to balance well both performance and annotation efficiency. The template method's predictions on the subjects it was trained with may provide good estimations of subsegmentations that could be used if labelling individual subjects is required, for example in MAS or deep learning methods, speeding the annotation process. The agreement observed between expert annotations and our automatic method supports this claim.

One method simplification could be introduced based on detecting the body-tail boundary using the pancreas segmentation centerline: the midpoint in length between the head-body boundary and the tip of the pancreatic tail would define the body-tail boundary more similarly to the anatomical definition used in this work, that is, “generally agreed to be located at the midpoint of the total length of the body and tail”, from [22]. We may also choose to fit each predicted boundary to a plane, similarly to the planes drawn in manual annotation, that is orthogonal to the pancreas centerline; in this scenario, the scalar distance between the manual boundary and predicted boundary planes may be used as the validation endpoint.

To date, we have studied regional differences for pancreatic PDFF, but note the method is suited to report differences in other biomarkers, such as T_1 , so long as the corresponding parametric maps are available within the imaging session.

4.1. Conclusion

This study demonstrated the feasibility of automated pancreas parts segmentation and downstream pancreatic imaging biomarker quantification by using groupwise registration of

whole-organ segmentations to a template, and subsequent annotation of the template image. This enables segmental characterization of heterogeneous pancreatic disease.

5. References

- [1] T. G. Cotter and M. Rinella, “Nonalcoholic Fatty Liver Disease 2020: The State of the Disease,” *Gastroenterology*, vol. 158, no. 7, pp. 1851–1864, May 2020, doi: 10.1053/J.GASTRO.2020.01.052.
- [2] A. Mathur *et al.*, “Nonalcoholic fatty pancreas disease,” *Hpb*, vol. 9, no. 4, pp. 312–318, 2007, doi: 10.1080/13651820701504157.
- [3] M. M. Smits and E. J. M. van Geenen, “The clinical significance of pancreatic steatosis,” *Nature Reviews Gastroenterology and Hepatology*, vol. 8, no. 3, pp. 169–177, 2011, doi: 10.1038/nrgastro.2011.4.
- [4] R. G. Singh, H. D. Yoon, L. M. Wu, J. Lu, L. D. Plank, and M. S. Petrov, “Ectopic fat accumulation in the pancreas and its clinical relevance: A systematic review, meta-analysis, and meta-regression,” *Metabolism: Clinical and Experimental*, vol. 69, pp. 1–13, 2017, doi: 10.1016/j.metabol.2016.12.012.
- [5] M. Takahashi, M. Hori, R. Ishigamori, M. Mutoh, T. Imai, and H. Nakagama, “Fatty pancreas: A possible risk factor for pancreatic cancer in animals and humans,” *Cancer Science*, vol. 109, no. 10, pp. 3013–3023, 2018, doi: 10.1111/cas.13766.
- [6] C. Acharya, S. Navina, and V. P. Singh, “Role of pancreatic fat in the outcomes of pancreatitis,” *Pancreatology*, vol. 14, no. 5, pp. 403–408, Sep. 2014, doi: 10.1016/j.pan.2014.06.004.
- [7] I. Lingvay *et al.*, “Noninvasive quantification of pancreatic fat in humans,” *Journal of Clinical Endocrinology and Metabolism*, vol. 94, no. 10, pp. 4070–4076, 2009, doi: 10.1210/jc.2009-0584.
- [8] M. E. Tushuizen *et al.*, “Pancreatic fat content and β -cell function in men with and without type 2 diabetes,” *Diabetes Care*, vol. 30, no. 11, pp. 2916–2921, Nov. 2007, doi: 10.2337/dc07-0326.
- [9] R. Wagner *et al.*, “Metabolic implications of pancreatic fat accumulation,” *Nature Reviews Endocrinology*, vol. 0123456789, 2021, doi: 10.1038/s41574-021-00573-3.
- [10] S. Matsumoto *et al.*, “Uneven fatty replacement of the pancreas: Evaluation with CT,” *Radiology*, vol. 194, no. 2, pp. 453–458, 1995, doi: 10.1148/radiology.194.2.7824726.

- [11] X. Wang *et al.*, “Regional Differences in Islet Distribution in the Human Pancreas - Preferential Beta-Cell Loss in the Head Region in Patients with Type 2 Diabetes,” *PLoS ONE*, vol. 8, no. 6, pp. 1–9, 2013, doi: 10.1371/journal.pone.0067454.
- [12] K. Suda, M. Takase, K. Takei, T. Kumasaka, and F. Suzuki, “Histopathologic and immunohistochemical studies on the mechanism of interlobular fibrosis of the pancreas,” *Archives of Pathology and Laboratory Medicine*, vol. 124, no. 9, pp. 1302–1305, 2000, doi: 10.5858/2000-124-1302-haisot.
- [13] I. Modolell, L. Guarner, and J. R. Malagelada, “Vagaries of clinical presentation of pancreatic and biliary tract cancer,” *Annals of Oncology*, vol. 10, pp. S82–S84, 1999, doi: 10.1093/annonc/10.suppl_4.S82.
- [14] D. P. Ryan, T. S. Hong, and N. Bardeesy, “Pancreatic adenocarcinoma,” *The New England journal of medicine*, vol. 371, no. 11, pp. 1039–1049, 2014, doi: 10.1056/NEJMra1404198.
- [15] T. Tirkes, C. Lin, E. L. Fogel, S. S. Sherman, Q. Wang, and K. Sandrasegaran, “ T_1 mapping for diagnosis of mild chronic pancreatitis,” *Journal of Magnetic Resonance Imaging*, vol. 45, no. 4, pp. 1171–1176, Apr. 2017, doi: 10.1002/jmri.25428.
- [16] H. H. Hu, H. W. Kim, K. S. Nayak, and M. I. Goran, “Comparison of fat-water MRI and single-voxel MRS in the assessment of hepatic and pancreatic fat fractions in humans,” *Obesity*, vol. 18, no. 4, pp. 841–847, 2010, doi: 10.1038/oby.2009.352.
- [17] A. Al-Mrabeh *et al.*, “2-year remission of type 2 diabetes and pancreas morphology: a post-hoc analysis of the DiRECT open-label, cluster-randomised trial,” *The Lancet Diabetes & Endocrinology*, vol. 8, no. 12, pp. 939–948, Dec. 2020, doi: 10.1016/S2213-8587(20)30303-X.
- [18] K. Sandrasegaran *et al.*, “Diffusion-weighted imaging in characterization of cystic pancreatic lesions,” *Clinical Radiology*, vol. 66, no. 9, pp. 808–814, Sep. 2011, doi: 10.1016/J.CRAD.2011.01.016.
- [19] M. Barral *et al.*, “Diffusion-weighted MR imaging of the pancreas: Current status and recommendations,” *Radiology*, vol. 274, no. 1, pp. 45–63, 2015, doi: 10.1148/radiol.14130778.
- [20] C. Nadarajah *et al.*, “Association of pancreatic fat content with type II diabetes mellitus,” *Clinical Radiology*, vol. 75, no. 1, pp. 51–56, Jan. 2020, doi: 10.1016/j.crad.2019.05.027.

- [21] J.-P. Kühn *et al.*, “Pancreatic Steatosis Demonstrated at MR Imaging in the General Population: Clinical Relevance,” *Radiology*, vol. 276, no. 1, pp. 129–136, Jul. 2015, doi: 10.1148/radiol.15140446.
- [22] K. Suda, B. Nobukawa, M. Takase, and T. Hayashi, “Pancreatic segmentation on an embryological and anatomical basis,” *Journal of Hepato-Biliary-Pancreatic Surgery*, vol. 13, no. 2, pp. 146–148, 2006, doi: 10.1007/s00534-005-1039-3.
- [23] T. Takada, H. Yasuda, K. Uchiyama, H. Hasegawa, T. Iwagaki, and Y. Yamakawa, “A proposed new pancreatic classification system according to segments: Operative procedure for a medial pancreatic segmentectomy,” *Journal of Hepato-Biliary-Pancreatic Surgery*, vol. 1, no. 3, pp. 322–325, 1994, doi: 10.1007/BF02391090.
- [24] A. Al-Mrabeh, K. G. Hollingsworth, S. Steven, D. Tiniakos, and R. Taylor, “Quantification of intrapancreatic fat in type 2 diabetes by MRI,” *PLoS ONE*, vol. 12, no. 4, pp. 1–19, 2017, doi: 10.1371/journal.pone.0174660.
- [25] S. Boga *et al.*, “Liver and pancreas: ‘Castor and Pollux’ regarding the relationship between hepatic steatosis and pancreas exocrine insufficiency,” *Pancreatology*, vol. 20, no. 5, pp. 880–886, 2020, doi: 10.1016/j.pan.2020.04.020.
- [26] T. J. Littlejohns *et al.*, “The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions,” *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020, doi: 10.1038/s41467-020-15948-9.
- [27] Y. Saisho, “Pancreas volume and fat deposition in diabetes and normal physiology: Consideration of the interplay between endocrine and exocrine pancreas,” *Review of Diabetic Studies*, vol. 13, no. 2–3, pp. 132–147, 2016, doi: 10.1900/RDS.2016.13.132.
- [28] J. P. Gilbeau, V. Poncet, E. Libon, G. Derue, and F. R. Heller, “The density, contour, and thickness of the pancreas in diabetics: CT findings in 57 patients,” *American Journal of Roentgenology*, vol. 159, no. 3, pp. 527–531, 1992, doi: 10.2214/ajr.159.3.1503017.
- [29] K. Karasawa *et al.*, “Multi-atlas pancreas segmentation: Atlas selection based on vessel structure,” *Medical Image Analysis*, vol. 39, pp. 18–28, Jul. 2017, doi: 10.1016/j.media.2017.03.006.
- [30] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.
- [31] J. Schlemper *et al.*, “Attention gated networks: Learning to leverage salient regions in medical images,” *Medical Image Analysis*, vol. 53, pp. 197–207, Apr. 2019, doi: 10.1016/J.MEDIA.2019.01.012.

- [32] G. Fontana *et al.*, “MRI quantification of pancreas motion as a function of patient setup for particle therapy -A preliminary study,” *Journal of Applied Clinical Medical Physics*, vol. 17, no. 5, pp. 60–75, 2016, doi: 10.1120/jacmp.v17i5.6236.
- [33] Z. Arya, G. Ridgway, A. Jandor, and P. Aljabar, *Deep Learning-Based Landmark Localisation in the Liver for Couinaud Segmentation*, vol. 12722 LNCS. Springer International Publishing, 2021. doi: 10.1007/978-3-030-80432-9_18.
- [34] M. Sinclair *et al.*, “Atlas-ISTN: Joint Segmentation, Registration and Atlas Construction with Image-and-Spatial Transformer Networks,” 2020.
- [35] J. Owler *et al.*, *Pancreas Volumetry in UK Biobank: Comparison of Models and Inference at Scale*, vol. 12722 LNCS. Springer International Publishing, 2021. doi: 10.1007/978-3-030-80432-9_21.
- [36] J. Ashburner and K. J. Friston, “Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation,” *NeuroImage*, vol. 55, no. 3, pp. 954–967, Apr. 2011, doi: 10.1016/j.neuroimage.2010.12.049.
- [37] P. A. Yushkevich *et al.*, “User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability,” *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006, doi: 10.1016/j.neuroimage.2006.01.015.
- [38] D. G. Altman and J. M. Bland, “Measurement in Medicine: The Analysis of Method Comparison Studies,” *The Statistician*, vol. 32, no. 3, p. 307, Sep. 1983, doi: 10.2307/2987937.
- [39] A. Triay Bagur, C. Hutton, B. Irving, M. L. Gyngell, M. D. Robson, and M. Brady, “Magnitude-intrinsic water–fat ambiguity can be resolved with multipeak fat modeling and a multipoint search method,” *Magnetic Resonance in Medicine*, vol. 82, no. 1, pp. 460–475, 2019, doi: 10.1002/mrm.27728.
- [40] G. Hamilton *et al.*, “In vivo characterization of the liver fat 1H MR spectrum,” *NMR in Biomedicine*, vol. 24, no. 7, pp. 784–790, 2011, doi: 10.1002/nbm.1622.

List of Figures and Tables

Figure 1. Whole-pancreas segmentations from the template construction ('training') dataset and the validation dataset, sorted by subjects' age (females in red, males in blue).

Figure 2. Method description. (top) Overall pipeline for whole-pancreas segmentation and parts segmentation. (bottom-left) (1) Offline groupwise registration of the whole-pancreas segmentation generated a population average ('template'), on which (2) the parts were manually annotated ('parts template', head: blue, body: green, tail: yellow). (bottom-right) For a new subject, the method (1) computes a registration transformation from the subject's segmentation to the template, (2) applies the inverse transformation on the parts template, and (3) propagates the warped parts template labels to the segmentation.

Figure 3. Pancreas segmentation (head: blue, body: green, tail: yellow) enables quantification of pancreas imaging biomarkers by parts, for example proton density fat fraction (PDFF), by intersection of the segmentation with the quantitative scan. PDFF, proton density fat fraction

Figure 4. Qualitative evaluation of parts segmentations from Rater 1, Rater 2, Rater 3, Rater 4, Automated k-means method, Automated groupwise registration method, respectively (head: blue, body: green, tail: yellow). The first 10 subjects of the validation set are shown.

Figure 5. Inter-observer variation of derived PDFF quantification from the manual experts' annotations.

Figure 6. Manual experts' annotations vs automated subsegmentations-derived PDFF quantification: differences by parts. 'Manual vs auto' comparisons are presented for both the k-means method and the groupwise registration method.

Figure 7. Example of regional pancreatic fat quantification for one subject in the validation set. Parts segmentation is shown on the PDFF image for Rater 1, Rater 2, Rater 3, Groupwise registration method, and K-means method, respectively (head: blue, body: green, tail: yellow). PDFF, proton density fat fraction

Figure 8. Pancreatic fat quantification by parts in groups from UK Biobank: (1) Type 2 diabetics (self-reported), (2) non-diabetics matched by gender, age, and BMI, (3) non-diabetics matched by gender and age and with low BMI. The number of segments available is displayed next to the part. BMI, body mass index

Table 1. Direct validation metrics, expert raters combined. Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (95%HD in mm) are reported as mean \pm standard deviation. Part volumes differences are reported in mL as bias [lower LoA, upper LoA]. Intra-observer agreement and inter-observer agreement are reported. LoA, limits of agreement

Table 2. Direct validation metrics, expert raters combined. Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (95%HD in mm) are reported as mean \pm standard deviation. Part volumes differences are reported in mL as bias [lower LoA, upper LoA]. Manual vs Automated agreement is reported for each of the existing automated methods. LoA, limits of agreement