Pancreas Volumetry in UK Biobank: Comparison of Models and Inference at Scale

James Owler[1(✉)], Alexandre Triay Bagur[1,2], Scott Marriage[1], Zobair Arya[1], Paul Aljabar[1], John McGonigle[1], Sir Michael Brady[1,3], and Daniel Bulte[2]

[1] Perspectum Ltd., Oxford, UK
james.owler@perspectum.com
[2] Department of Engineering Science, University of Oxford, Oxford, UK
[3] Department of Oncology, University of Oxford, Oxford, UK

Abstract. The UK Biobank imaging sub-study enables large-scale measurement of pancreas volume, an important biomarker in metabolic disease, including diabetes. Previous methods utilised a pancreas-specific (PS) 3D MRI UK Biobank acquisition to automatically measure pancreas volume. This may lead to a clinically significant underestimation of volume, due to partial coverage of the pancreas in these acquisitions. To address this, we propose a pipeline for the accurate measurement of pancreas volume using stitched whole-body (WB) 3D MRI UK Biobank acquisitions and deep learning-based segmentation. We implement and compare the performance of six different U-Net-like model architectures, leveraging attention layers, recurrent layers, and residual blocks. Furthermore, we investigate pancreas volumetry in 42,313 subjects, separated by sex, and present novel results concerning the change in pancreas volume throughout the course of a day (diurnal variation). To the best of our knowledge, this is the largest pancreas volumetry study to date and the first to propose a pipeline using the whole-body UK Biobank MRI acquisitions to measure pancreas volume.

Keywords: Pancreas segmentation · Deep learning · UK Biobank

1 Introduction

Pancreas volume has been shown to change with age and in diseases such as pancreatitis, type 1 diabetes, and type 2 diabetes [1–5]. Pancreas volume is typically measured following segmentation. Manual pancreas segmentation is labour-intensive and the delineation of a three-dimensional shape, with ill-defined boundaries extending across multiple views, is prone to substantial inter-rater and intra-rater variability. Automating pancreas segmentation can alleviate these problems by increasing reproducibility and decreasing subjectivity. Automation can also allow for large-scale model deployment, leading to practical implementations for clinical decision-making and population health research.

B. W. Papież et al. (Eds.): MIUA 2021, LNCS 12722, pp. 265–279, 2021.
https://doi.org/10.1007/978-3-030-80432-9_21

Deep learning-based pancreas segmentation methods have been proposed to automate pancreas segmentation [6,7].

UK Biobank is one of the largest resources of imaging and non-imaging medical data in the world. This resource opens up the possibility of exploratory research into the realm of precision medicine [8]. The UK Biobank imaging substudy aims to scan a total of 100,000 volunteers at multiple timepoints [9]; part of this sub-study includes numerous MRI imaging acquisitions. One such acquisition is a dedicated pancreas volumetric interpolated breath-hold examination (VIBE). Previous works have used this pancreas-specific scan for pancreas segmentation, volumetry estimation, shape measurement and downstream pancreatic quantification [10,11]. The high resolution nature of the pancreas-specific scan allows models to more easily learn useful representations of the pancreas and to better quantify biomarkers like surface lobularity [5,12]. However, in many cases, there is only partial coverage of the pancreas, often missing parts of the pancreas head region (Fig. 1). The longitudinal nature of UK Biobank motivates the need for accurate and precise pancreas volume measurement, in order to detect small (clinically) meaningful changes caused by aging and pathological processes. The partial coverage effect observed in pancreas-specific images could lead to significant inaccuracies in volume measurements, hindering the ability to detect such small changes. For reference, a study by Saisho et al. [2] reported a 74.9 cm^3 total pancreas volume in type 2 diabetics and 70.0 cm^3 volume in age-, sex-, and BMI-matched controls.

The UK Biobank imaging protocol also contains a whole-body (neck-to-knee) 2-point Dixon acquisition, acquired sequentially in multiple breath-hold volumes. Although they have been acquired at slightly lower resolution than the pancreas-specific scan, they have the advantage that they contain full coverage of the pancreas. To the best of our knowledge, no study investigating pancreas volumetry in UK Biobank has used these whole-body acquisitions to measure pancreas volume. These whole-body scans can provide accurate and consistent volume measurements across the UK Biobank population. In addition, they can provide insight into interrelations with other imaging (e.g. quantitative MRI, volumetry from other organs) and non-imaging (e.g. blood tests, diagnoses, genetics) biomarkers, in order to better understand and improve treatment of disease.

In this work, we propose a novel pipeline to study pancreas volumetry in UK Biobank. This pipeline can be easily extended to study other organs captured in the whole-body acquisition. The pipeline includes stitching, registration to a common coordinate space, cropping out the abdominal region, deep learning-based segmentation model training, testing and prediction. We leverage a previously presented pancreas segmentation model [11] to build a large annotated dataset of whole-body pancreas segmentations. We implement six different U-Net [13] based models that have been proposed in the literature, some with existing pancreas segmentation applications [14], and compared their performance. Finally, we estimate the pancreas volume under-estimation caused by the partial coverage in the pancreas-specific acquisitions and investigate pancreas volume population metrics, including diurnal variation of pancreas volume.

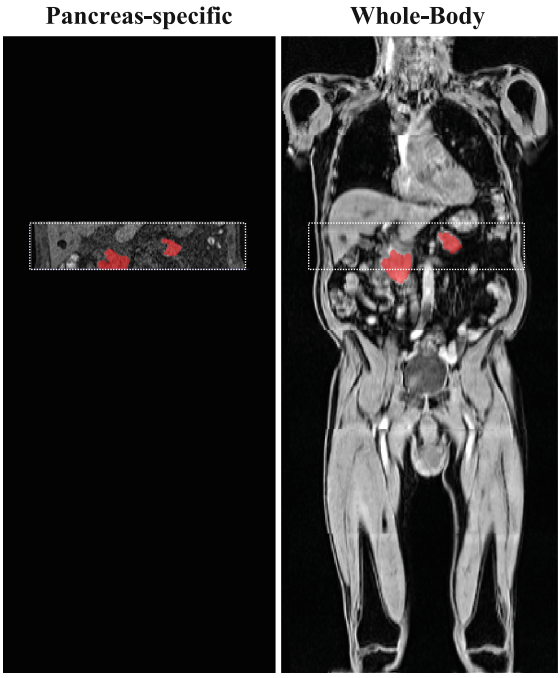

Fig. 1. A coronal slice through the 3D pancreas-specific scan (left), overlayed on a coronal slice through the 3D whole-body scan (right). The bottom of the pancreas has been cropped out in the PS scan which has led to the underestimation of pancreas volume. The measured pancreas volume from the PS scan was 78.6 cm^3, whereas the measured pancreas volume from the whole-body scan was 91.9 cm^3 – a 14.4 % underestimation.

2 Materials and Methods

2.1 Data Acquisition

Imaging data was acquired with a Siemens Aera 1.5 T (Siemens Healthineers AG, Erlangen, Germany). "Pancreas fat - DICOM" (Data-Field ID 20202 in the UK Biobank Showcase[1]) volumetric acquisition, which we are referring to here as the pancreas-specific (PS) scan, targets the abdominal location of the pancreas. The PS scan used the FLASH-3D acquisition (echo time (TE)/repetition time (TR) = 1.15/3.11 ms, voxel size = 1.1875 × 1.1875 × 1.6 mm), with 10° flip angle and fat suppression.

"Dixon technique for internal fat - DICOM" (Data-Field ID 20201) volumetric acquisition, which we are referring to here as the whole-body (WB) scan, involved multiple dual echo Dixon VIBE volumes, acquired over the course of 6 min to provide water/fat separated overlapping volumes from the neck to the knees. Further details about the acquisition can in found in [15].

[1] https://biobank.ndph.ox.ac.uk/showcase/browse.cgi.

2.2 Data Labelling and Preprocessing

First, overlapping volumes from the WB acquisition were automatically stitched together, following a round of N4 bias field correction [16] on each volume block. To deal with an MRI wrapping artefact present in the data, the top two and the bottom two slices from each block were removed prior to stitching. Linear normalisation was then applied across blocks, to correct for any cross-block intensity inhomogeneity. This resulted in a relatively 'clean' stitched image, with a homogeneous intensity throughout.

Second, in order to build an annotated dataset for model training, we leveraged a model previously trained on PS acquisition images that were labelled by an expert [11]. Labels, obtained from the PS model were propagated across to the WB images using DICOM header geometry information. 200 WB images, along with their propagated label counterpart, were selected by visual inspection. Subjects that had minimal movement between the two types of acquisitions and thus, whose propagated labels best aligned with the WB image were selected. Corrections, including the addition of the missing pancreas information, were made after the alignment. Corrections were performed using the 3D brush tool in ITK-SNAP [17]. This labelling process allowed us to quickly build an annotated dataset.

Third, a representative WB image was selected to provide a dedicated 'reference' coordinate space to which all other WB images were aligned via affine image registration. Here, we used the affine registration implementation from ANTS [18]. This enabled us to heuristically crop out the abdominal region from the WB image, resulting in a consistent image size and similar coverage of abdominal organs across all subjects. As we are only interested in the pancreas region, training a model using the full WB image would have been inefficient in terms of GPU memory useage, in addition to potentially degrading the performance of the model. The same registration and cropping was applied to the WB pancreas label. Segmentations were transformed back to their original coordinate space, via the inverse affine transformation, when calculating a final measurement of pancreas volume.

Lastly, we clipped image intensities at the 99th percentile value, normalised voxel values between 0 and 1, and randomly split the dataset as follows: 70% training, 20% testing, and 10% validation.

2.3 Model Architectures

We compared the performance of six different semantic segmentation networks based on previously reported architectures. The first of these was a conventional 3D U-Net with an encoder path, decoder path and skip connections [19]. The details of this network can be found in Figs. 2a and 3a.

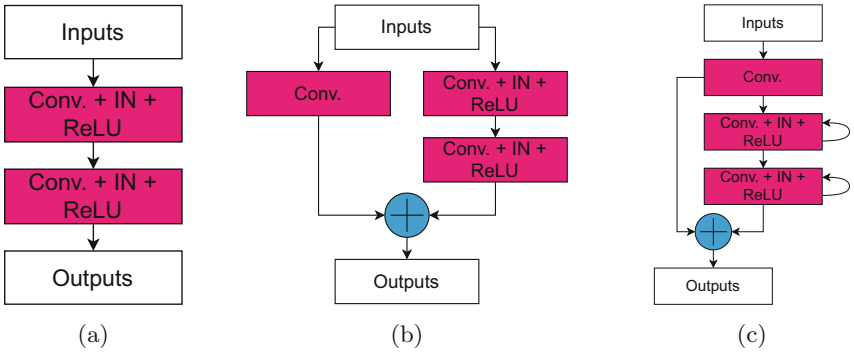

Fig. 2. Different variants of convolutional blocks. (a) Conventional convolutional block, (b) Residual convolutional block, (c) Recurrent residual convolutional block.

The second network was an attention U-Net (AU-Net) as reported by Oktay et al. [14]. The idea behind AU-Nets is to include a mechanism such that the network learns to focus only on the features of an image that are relevant to the task. Given that the pancreas encompasses only a small part of the cropped whole-body image, AU-Nets have potential to be suitable for our application. We implemented the AU-Net by, prior to the concatenation step in each skip connection, 'gating' the incoming feature maps from the encoder layer as shown in Fig. 3b. This gating step is as follows: the n_x feature maps from the encoder layer are denoted by x, and the corresponding n_g feature maps from the decoder layer that are typically concatenated with x in a skip connection are denoted by g. First, x is downsampled to the same spatial resolution as g. A set of attention weights, α, is then obtained through the following operation: $\alpha = \sigma_2(\psi^T(\sigma_1(W_x^T x + W_g^T g + b_1)) + b_2)$. Here, W_x and W_g represent $1 \times 1 \times 1$ convolution operations that output n_x features, σ_1 is a ReLu activation function, ψ is a $1 \times 1 \times 1$ convolution that outputs 1 feature map, σ_2 is a sigmoid activation function and the b vectors are bias terms. Finally, α is resampled to the same spatial resolution as x, and x is multiplied by α to give the attention gated features $\hat{x}$.

The third architecture we investigated was a residual U-Net (RU-Net) as reported by Alom et al. [20]. Figure 2a shows that our conventional U-Net architecture consists of blocks that have two convolutional + instance-normalisation + ReLu (Conv + IN + ReLu) layers. In the RU-Net, these blocks are altered so that output of the second Conv + IN + ReLu layer is summed with a feature set based on the input to the first Conv + IN + ReLu layer, as shown in Fig. 2b. Note that the 'Conv' layer in Fig. 2b is a $1 \times 1 \times 1$ convolutional layer in the residual path to ensure that the dimensions of the two feature sets to be summed match. The motivation behind RU-Nets is that the residual paths can improve optimisation during training.

The fourth architecture that we investigated, also reported by Alom et al. [20], was a recurrent residual U-Net (R2U-Net). It builds on top of

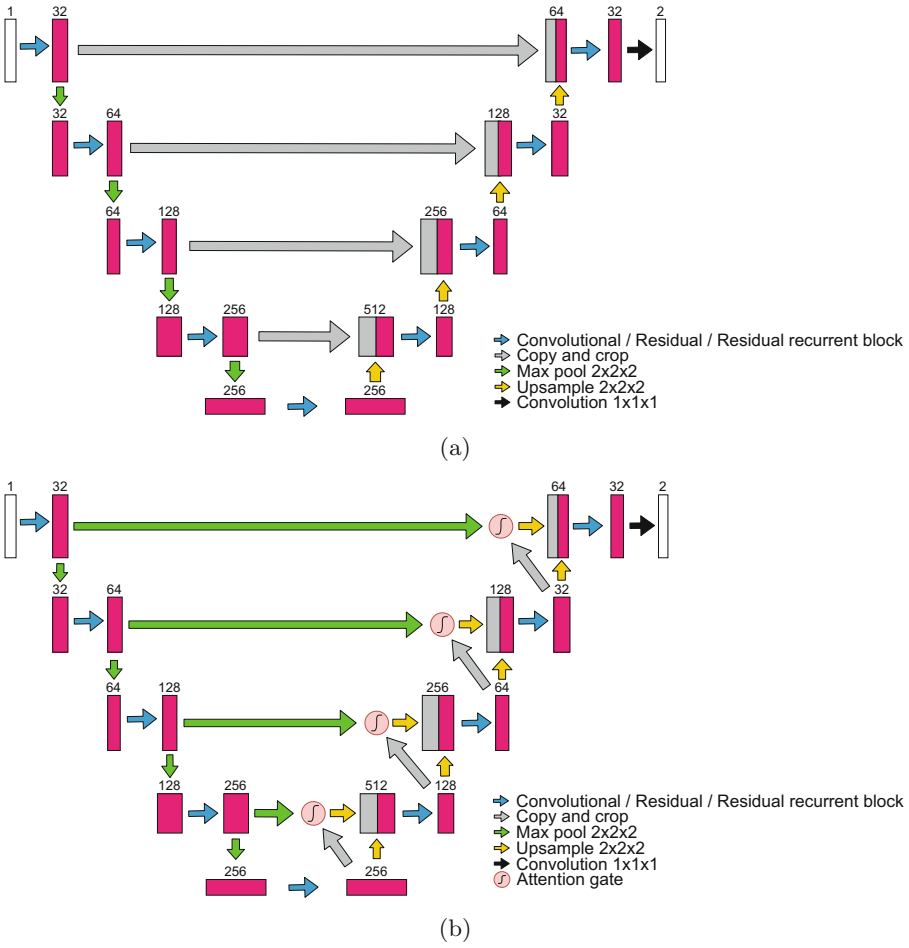

Fig. 3. The base U-Net model architectures for (a) a conventional U-Net and (b) an attention U-Net. Note that the blue arrows represent a block of operations, where the details of the block depend on the specific U-Net variant being implemented. The details for each type of block are shown in Fig. 2.

RU-Net by using recurrent convolutional layers (RCL), as opposed to conventional convolutional layers. R2U-Net has been shown to improve performance when compared to a conventional U-Net. In an RCL, features are 'accumulated' in a layer by forward propagating through the layer multiple times, and taking into account the feature maps from the previous propagations. For example, if x is the original input to the RCL, then the output of the RCL can be calculated by $o_t = W_f^T z_t + b$, where W_f represents a convolution, $z_1 = x$ for $t = 1$ and $z_t = o_{t-1} + z_{t-1}$ for $t > 1$. The output at each time-step is fed through a ReLu activation. We implemented the R2U-Net with two time-steps. The details of the

R2U-Net block is shown in Fig. 2c. As before, a 'Conv' layer is used to ensure that the dimensions match for any summation of features.

The U-Net variants above are modular, meaning they can be further combined with one another to utilise each of their potential benefits. Consequently, we investigated two further variants. The first was a residual U-Net with attention-gating (RAU-Net), the second was a recurrent residual U-Net with attention-gating (R2AU-Net).

2.4 Model Training and Testing

Before training, weights in each network were initialised using Glorot uniform initialisation [21]. Data augmentation and random shuffling were performed 'on-the-fly'. Augmentation included random rotations, within a range of -10° to +10° about the inferior-superior axis. We also randomly scaled the image size to between 95% and 105% of the original image dimensions. Zero-padding was used to keep to a consistent input dimension. Each network was trained for 100 epochs, with a batch size of 1 and a learning rate of 0.0005. Model weights were saved each time there was a decrease in validation loss. ADAM [22] optimisation was used, with a combined cross-entropy and soft dice loss function as shown in the following equation:

$$L = \frac{1}{N}\frac{1}{K}\sum_{i}^{N}\sum_{c}^{K}\left(1 - \frac{2\sum_{j}^{M} p_{icj}y_{icj} + \delta}{\sum_{j}^{M}(p_{icj} + y_{icj}) + \delta} - \frac{1}{M}\sum_{j}^{M} y_{icj}\log p_{icj}\right) \quad (1)$$

where $i = [1..N]$ is the subject index, $c = [1..K]$ is the class index, $j = [1..M]$ is the voxel index, p_{icj} is the predicted probability of voxel j for subject i belonging to class c, y_{icj} is a binary variable equal to 1 if the ground truth class for voxel j in subject i is c and 0 otherwise, and $\delta = 0.01$.

The models, training pipeline, and testing pipeline were all implemented using PyTorch[2]. Each model took approximately 5 h to train on an NVIDIA Tesla V100 GPU.

During inference, final pancreas segmentations were obtained by thresholding the label probabilities at 0.5, a threshold that selects the class label with the largest posterior probability under the assumption of equal misclassification costs. These segmentation labels were then resampled back to the original image size, using nearest-neighbour interpolation. Pancreatic volumes were calculated from the final segmentation labels.

To evaluate each model we used the commonly reported Dice Similarity Coefficient (DSC) and the 95th percentile of the Hausdorff Distance metric (HD95). Results from this evaluation can be seen in Sect. 3.1.

[2] https://pytorch.org/.

2.5 Model Inference at Scale

After evaluating the performance of each model variant, we used AU-Net to automatically segment 42,313 pancreas volumes in UK Biobank. The stitching, cropping, and pre-processing of each WB image was the same as described in Sect. 2.2. By utilising Terraform[3] to orchestrate multiple Amazon Web Services EC2 instances, we were able to obtain all 42,313 pancreas segmentations, and their corresponding volumes, in less than 4 h.

Pancreas Volume Within the UK Biobank Population. Using these automatically obtained pancreas segmentations, we compared volume measurements from the proposed WB segmentation method with volume measurements derived from the previously proposed PS segmentation model [11]. We used this comparison to estimate the extent of pancreas volume under-estimation in the PS scans. We also calculated the average pancreas volume for males, females, and combined males and females. These results are presented in Sect. 3.2 and 3.3, respectively.

Pancreas Volume Diurnal Variation. As an applied use-case of automated pancreas volumetry, and to demonstrate the versatility of large-scale research resources such as UK Biobank, we further investigated the natural change in pancreas volume (if any) throughout the course of the day. In UK Biobank, each subject is scanned just once, at an imaging session typically between the times of 9 am and 7 pm. Here, we rounded the timepoint at which a subject was scanned to the nearest hour; subjects scanned after 7:30 pm and before 8:30am were excluded to keep the number of subjects in each sub-group to greater than 1000. This resulted in 11 unique groups of pancreas volumes (mean n = 3791, range n = 2796–4194). We then calculated the median pancreas volume at each timepoint and observed the change in median volume between those timepoints. The sheer scale of UK Biobank allows us to investigate average changes in the human body, with the noise present from individual measurements largely mitigated. Using the whole-body images to measure pancreas volume also mitigates added noise from partial coverage in the PS acquisitions. One could measure the volume of the pancreas for the same individual at multiple timepoints throughout the day; however, it can be argued that observing the same average phenomena in tens of thousands of people provides greater validity to the result. Pancreas volume diurnal variation results are presented in Sect. 3.4.

3 Results

3.1 Model Evaluation

Qualitative Results. Figure 4 shows a qualitative comparison of the different models. Although it is difficult to draw any firm conclusions from a qualitative

[3] https://www.terraform.io/.

evaluation of predicted segmentations, one noticeable observation is that each of the automated models in Fig. 4 struggle to segment the pancreas towards the lowest extent of the organ. See Sect. 4 for further discussion.

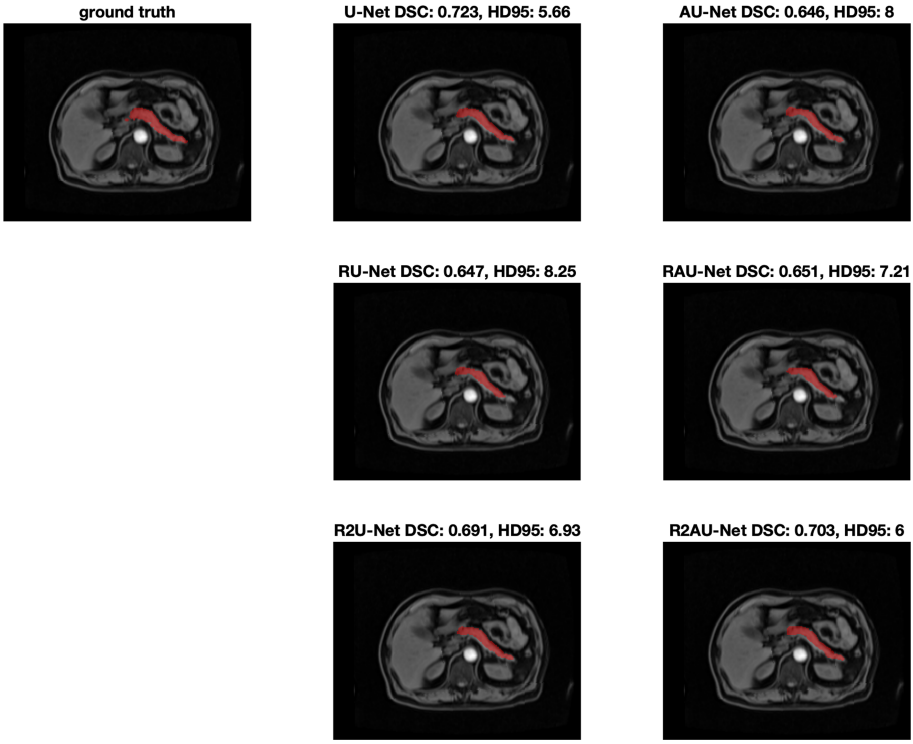

Fig. 4. Axial slice through a cropped WB image, showing automated pancreas segmentations and their respective performance metrics, for each type of model being compared.

Quantitative Results. Table 1 shows the DSC and HD95 metrics for the various models we investigated. Figures 5a and 5b also show boxplots for the these metrics. The conventional U-Net and R2U-Net models both resulted in the highest mean DSC, while the conventional U-Net resulted in the highest median DSC. When considering the standard deviation of DSC, it was R2U-Net that appeared to be least susceptible to outliers, with AU-Net in close second. In terms of HD95, AU-Net resulted in the lowest mean score and tightest standard

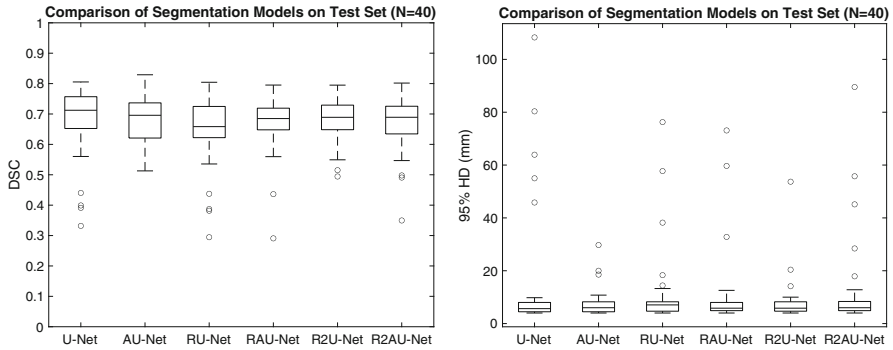

Fig. 5. Boxplots of the (a) DSC and (b) HD95 metrics for each type of model we investigated.

deviation. In general, no one model appeared to be clearly superior. The performances were similar across the board. Based on the fact that it was one of the models least susceptible to outliers, we decided to use AU-Net to derive the volumes presented in future sections.

Table 1. Model evaluation results. DSC - dice score; HD 95 - 95 percentile Hausdorff distance; SD - standard deviation.

Model	DSC			HD95		
	Median	Mean	SD	Median	Mean	SD
U-Net	0.712	**0.681**	0.114	5.66	13.7	23.2
AU-Net	0.696	0.675	0.078	6.00	**7.54**	**4.94**
RU-Net	0.658	0.652	0.112	7.07	10.1	14.4
RAU-Net	0.685	0.668	0.090	5.83	10.2	13.9
R2U-Net	0.689	**0.681**	**0.070**	5.83	7.87	8.04
R2AU-Net	0.689	0.670	0.088	6.00	11.29	16.51

3.2 Comparison with Volumetry from Pancreas-Specific Scan

Figure 6 shows the histogram of differences between volumes derived from the PS and the WB scans. We observed a mean difference of 11.7 cm^3, or 14.25%, ($p = 1.4 \times 10^{-288}$) between the two approaches.

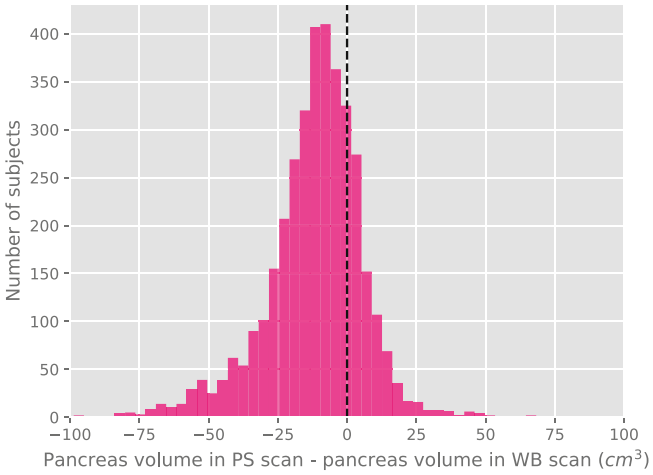

Fig. 6. Comparison of measured pancreas volume from the pancreas-specific scan segmentation model and the proposed whole-body scan segmentation model. A mean volume difference of 11.7 cm^3 (14.25%) was observed. n = 3672.

3.3 UK Biobank Population Volumetry

Table 2 shows median pancreas volume in UK Biobank for both males, females, and combined males and females. Figure 7 shows histograms of pancreas volumes for males and females. These results show a 13% difference between the average volume of the pancreas in males when compared with females.

Table 2. Average pancreas volumes in UK Biobank for males, females, and combined males and females.

	Number Quantified	Median (cm^3)	SD (cm^3)
Male	20395	70.0	19.8
Female	21918	62.0	17.1
Combined	42313	65.5	18.9

3.4 Pancreas Volume Diurnal Variation.

Fig. 8 shows that there is a marked variation in the volume of the pancreas throughout the course of a day. The largest change in total pancreatic volume of a 5.73% reduction ($p = 6.80 \times 10^{-20}$) was observed between the hours of 9am and 3pm. The largest change over the course of an hour, observed between 10 am and 11 am, was a reduction of 2.57% ($p=1.01\times10^{-5}$).

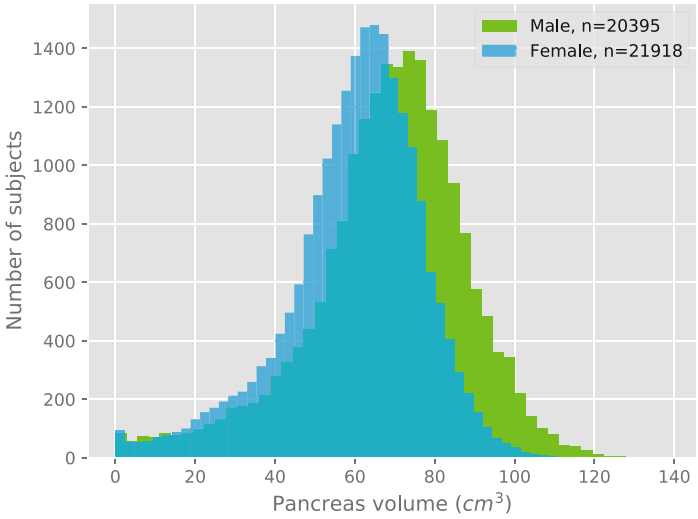

Fig. 7. Histograms of pancreas volume in UK Biobank for males and females. n = 42313.

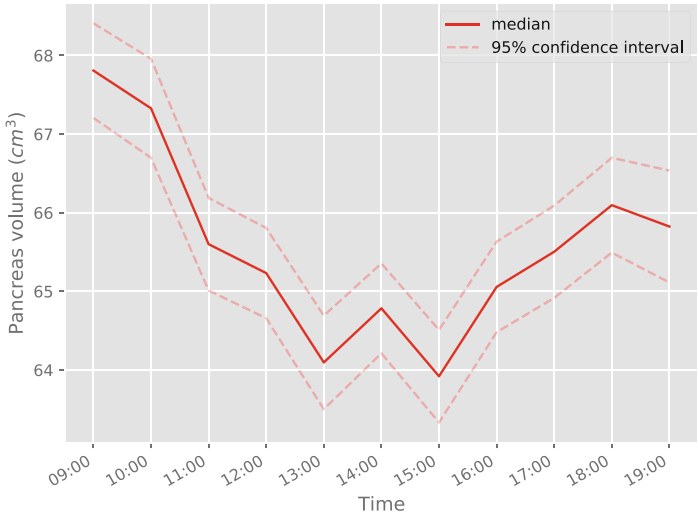

Fig. 8. Pancreas volume diurnal variation in UK Biobank (combined male and female). n = 41704.

4 Discussion and Conclusion

Pancreas volumetry in UK Biobank measured using the proposed pipeline, notably segmenting the pancreas on the WB acquisitions, agrees with reported values for nominally healthy populations [2] (N = 1,721). On the other hand,

pancreas volumetry performed using the PS acquisitions [10,11], underestimated volume by an average of 11.7 cm^3 compared to our method. To the best of our knowledge, this work is the largest attempt at accurate pancreas volume measurement in a nominally healthy population, which may be used for reference in future studies of age and disease.

Using the PS segmentation model allowed us to exploit the prior expert knowledge distilled in the annotations used for training the original model. It also enabled cheap, fast generation of good-quality starting estimates of pancreas labels for the WB images. One limitation of this approach is that, while these starting estimates were manually corrected when necessary, the selection of a subset of 'good' starting candidates from a larger dataset could have biased our dataset for WB model training. For instance, if we consider the extreme case of selecting a subset where no annotations are needed, the dataset might become limited towards those small pancreata that already fit in the PS scan volume, though that could be partly addressed using data augmentation.

We chose AU-Net to run at scale as it was less susceptible to outliers in both DSC and HD95. The differences in performance between the segmentation models were not found to be significant (when using a paired t-test); however, the test set was a relatively small sample. The 'effect size' of differences, in which it is difficult to gain insight to with a paired t-test alone, could be investigated further with more sophisticated Bayesian testing [23].

Although the performance of all the models presented here are on-par with other state-of-the-art pancreas segmentation methods [24], there is scope for improvement. Due to all of the cropped UK Biobank whole-body images being the same resolution, a patch-based segmentation approach could improve segmentation performance. This would mean that neither the input image or the output label would require any resampling, thereby avoiding resampling errors at object boundaries. This type of resampling error can be particularly detrimental in smaller organs, such as the pancreas, when using methods like nearest neighbour interpolation. This resampling error could also lead to inaccuracies in pancreas surface lobularity measures.

In terms of the diurnal variation of pancreas volume, we are not aware of the biological mechanism that causes this change; however, a similar pattern has been observed in other organs in the body [25]. It is important to note that there is a marked change in pancreas volume throughout the day, which should be considered when making clinical decisions based on volume assessments. This change could be corrected for via normalisation, although more experimentation is needed to tease out any unforeseen biases in the data before presenting any correction methods.

In conclusion, we have highlighted clinically significant underestimation of pancreas volume in UK Biobank, caused by partial coverage in the pancreas-specific acquisition. We presented a comparison of 6 different variants of U-Net models for pancreas segmentation in whole-body MRI. We also proposed a pipeline for efficient data labelling, using a previously trained PS model, and deployment of a trained model on a large scale. We believe the culmination of

large data sources, such as UK Biobank, with deep learning methods, and cloud computing has exciting potential to provide a better insight into population health, allow for the exploration of novel biomarkers, and improve patient care.

Acknowledgements. We would like to acknowledge Perspectum Ltd and the Engineering and Physical Sciences Research Council (EPSRC) for funding and support. This research has been conducted using the UK Biobank resource under application 9914.

References

1. Schrader, H., et al.: Reduced pancreatic volume and β-cell area in patients with chronic pancreatitis. Gastroenterology **136**(2), 513–522 (2009). http://dx.doi.org/10.1053/j.gastro.2008.10.083
2. Saisho, Y., et al.: Pancreas volumes in humans from birth to age one hundred taking into account sex, obesity, and presence of type-2 diabetes. Clin. Anat. **20**(8), 933–942 (2007)
3. Saisho, Y.: Pancreas volume and fat deposition in diabetes and normal physiology: consideration of the interplay between endocrine and exocrine pancreas. Rev. Diabet. Stud. **13**(2–3), 132–147 (2016)
4. Macauley, M., Percival, K., Thelwall, P.E., Hollingsworth, K.G., Taylor, R.: Altered volume, morphology and composition of the pancreas in type 2 diabetes. PLoS ONE **10**(5), 1–14 (2015)
5. Al-Mrabeh, A., et al.: 2-year remission of type 2 diabetes and pancreas morphology: a post-hoc analysis of the DiRECT open-label, cluster-randomised trial. Lancet Diabetes Endocrinol. **8**(12), 939–948 (2020). http://dx.doi.org/10.1016/S2213-8587(20)30303-X
6. Cai, J., Lu, L., Xing, F., Yang, L.: Pancreas segmentation in CT and MRI via task-specific network design and recurrent neural contextual learning. In: Lu, L., Wang, X., Carneiro, G., Yang, L. (eds.) Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. ACVPR, pp. 3–21. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13969-8_1
7. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021). http://dx.doi.org/10.1038/s41592-020-01008-z
8. Sudlow, C., et al.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. Plos Med **12**(3), e1001779 (2015)
9. Littlejohns, T.J., et al.: The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature Commun. **11**(1), 1–12 (2020). http://dx.doi.org/10.1038/s41467-020-15948-9
10. Liu, Y., et al.: Genetic architecture of 11 abdominal organ traits derived from abdominal MRI using deep learning, pp. 1–66 (2020)
11. Bagur, A.T., Ridgway, G., McGonigle, J., Brady, S.M., Bulte, D.: Pancreas segmentation-derived biomarkers: volume and shape metrics in the UK biobank imaging study. In: Papież, B.W., Namburete, A.I.L., Yaqub, M., Noble, J.A. (eds.) MIUA 2020. CCIS, vol. 1248, pp. 131–142. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52791-4_11

12. Calandra, A., Sartoris, R., Lee, K.J., Gauss, T., Vilgrain, V., Ronot, M.: Quantification of pancreas surface Lobularity on CT: a feasibility study in the normal pancreas (2020)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
15. Linge, J., et al.: Body composition profiling in the UK biobank imaging study. Obesity **26**(11), 1785–1795 (2018)
16. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010). https://www.ncbi.nlm.nih.gov/pubmed/20378467, www.ncbi.nlm.nih.gov/pmc/PMC3071855/
17. Yushkevich, P.A., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage **31**(3), 1116–1128 (2006)
18. Avants, B.B., Tustison, N., Song, G.: Advanced normalization tools (ants). Insight J. **2**(365), 1–35 (2009)
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, pp. 565–571 (2016)
20. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. J. Med. Imaging **6**(1), 014006 (2019)
21. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. J. Mach. Learn. Res. **9**, 249–256 (2010)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J. Mach. Learn. Res. **18**, 1–36 (2017)
24. Heinrich, M.P., Oktay, O., Bouteldja, N.: OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. Med. Image Anal. **54**, 1–9 (2019)
25. Owler, J., McGonigle, J., Robson, M., Brady, M., Banerjee, R.: Liver volume diurnal variation in UK biobank. In: The Liver Meeting Digital Experience™. AASLD (2020)