Journal of the American Medical Informatics Association, 28(6), 2021, 1178–1187
doi: 10.1093/jamia/ocaa342
Advance Access Publication Date: 12 February 2021
Research and Applications

Research and Applications

Quantification of abdominal fat from computed tomography using deep learning and its association with electronic health records in an academic biobank

Matthew T. MacLean,[1,2] Qasim Jehangir,[3] Marijana Vujkovic,[3] Yi-An Ko,[1] Harold Litt,[1] Arijitt Borthakur,[1] Hersh Sagreiya,[1] Mark Rosen,[1] David A. Mankoff,[1] Mitchell D. Schnall,[2] Haochang Shou,[4] Julio Chirinos,[3] Scott M. Damrauer,[5] Drew A. Torigian,[1] Rotonya Carr,[3] Daniel J. Rader,[2,3,*] and Walter R. Witschey ,[1,*]

[1]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [3]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [4]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and [5]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

*These authors contributed equally to this work.

Corresponding Author: Walter R. Witschey, Department of Radiology, Perelman Center for Advanced Medicine, 3400 Civic Center Boulevard, Philadelphia, PA 19104, USA witschey@pennmedicine.upenn.edu

Received 24 August 2020; Revised 30 November 2020; Editorial Decision 19 December 2020; Accepted 6 January 2021

ABSTRACT

Objective: The objective was to develop a fully automated algorithm for abdominal fat segmentation and to deploy this method at scale in an academic biobank.
Materials and Methods: We built a fully automated image curation and labeling technique using deep learning and distributive computing to identify subcutaneous and visceral abdominal fat compartments from 52,844 computed tomography scans in 13,502 patients in the Penn Medicine Biobank (PMBB). A classification network identified the inferior and superior borders of the abdomen, and a segmentation network differentiated visceral and subcutaneous fat. Following technical evaluation of our method, we conducted studies to validate known relationships with visceral and subcutaneous fat.
Results: When compared with 100 manually annotated cases, the classification network was on average within one 5-mm slice for both the superior (0.4 ± 1.1 slice) and inferior (0.4 ± 0.6 slice) borders. The segmentation network also demonstrated excellent performance with intraclass correlation coefficients of 1.00 ($P < 2 \times 10^{-16}$) for subcutaneous and 1.00 ($P < 2 \times 10^{-16}$) for visceral fat on 100 testing cases. We performed integrative analyses of abdominal fat with the phenome extracted from the electronic health record and found highly significant associations with diabetes mellitus, hypertension, and renal failure, among other phenotypes.
Conclusions: This work presents a fully automated and highly accurate method for the quantification of abdominal fat that can be applied to routine clinical imaging studies to fuel translational scientific discovery.

Key words: deep learning, subcutaneous fat, machine learning, Penn Medicine biobank, visceral fat, body mass index

© The Author(s) 2021. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For permissions, please email: journals.permissions@oup.com
This article is published and distributed under the terms of the Oxford University Press, Standard Journals Publication Model (https://academic.oup.com/journals/pages/open_access/funder_policies/chorus/standard_publication_model)

INTRODUCTION

Medical centers collect enormous quantities of imaging data that could be extremely valuable for translational science, but many quantitative traits are not systematically extracted. One of these quantitative traits is abdominal adipose tissue volume, which is highly relevant to human health and disease and can be quantified from medical images such as computed tomography (CT) scans. Obesity, a condition of increased adipose tissue, has been associated with numerous diseases including cardiovascular disease, diabetes, stroke, and cancer.[1,2] However, obesity is diagnosed by body mass index (BMI), which is a poor measure of fat, as it is calculated using only weight and height and does not account for variations in body composition.[3]

Furthermore, not all fat is equal. Visceral adipose tissue (VAT), which is located within the abdominal cavity adjacent to vital organs, portends an even greater risk for many pathologies including cardiovascular disease, insulin resistance, and certain cancers.[4] However, the complex relationship between VAT and disease is not yet understood.[4–7] Fundamental to the continued study of this topic is the ability to accurately quantify and distinguish VAT and subcutaneous adipose tissue (SAT). Many studies rely on waist circumference as a proxy for VAT.[8–10] However, while waist circumference has demonstrated clinical significance beyond BMI, it correlates more strongly with SAT than VAT.[11,12] CT scans offer a solution as they are routinely performed and provide a cross-sectional view of anatomy that is often used to measure abdominal adipose tissue.

Robust and automatic techniques for extracting fat imaging traits from CT could help in the at-scale task of processing data in large biobank populations, developing precision medicine algorithms, expediting clinical workflows, or even deepening our understanding of machine learning bias. There are numerous biobanks in the United States and internationally that collect genetic data and correlate this with electronic health record (EHR)–documented pathology.[13] By better understanding the relationship between body composition and genetics, environment, and disease, we can begin to offer patients more targeted, precision medicine interventions. This knowledge combined with the processing algorithm could then be integrated into a radiology practice to provide valuable information to clinicians making care decisions. Furthermore, machine learning algorithms excel in pattern recognition but sometimes make predictions that are incorrect and based on biases learned during training.[14,15] These biases are easiest to detect when an algorithm is applied to a large diverse cohort, but this requires a method that can be efficiently applied at scale.

Many methods that have been proposed for automatic fat quantification have limitations that prevent their application to a large clinical cohort. These methods often rely on expected anatomic profiles to apply a statistical model.[16–24] The active contour model is an example of a common segmentation technique that has been applied to body composition analysis,[20,24] and works by minimizing an energy function designed to create a smooth boundary at regions of high gradients.[25] However, this approach can easily fail when image noise creates local minima or when the object has boundaries with high curvature.[26] Another approach is model-based segmentation in which a model is constructed based on expected geometry and then deformed to identify the object of interest on new cases.[21,23] This approach can work well with homogeneous data, but variations in shape and size limit its generalizability.[27] In summary, the previously highlighted methods rely on expected attenuation profiles or geometric properties and can work well on curated datasets but easily fail when artifacts, anatomical variation, or unexpected pathology are encountered.

Deep learning offers a data-driven approach to overcome these limitations by learning and prioritizing features based on training data. Deep learning has been applied to a range of biomedical segmentation and classification tasks with impressive results.[29–32] Specific to abdominal fat quantification, deep learning has been utilized in multiple studies.[28,33–35] However, many of these methods are only applicable to single-slice quantification and do not address identifying the slices of interest. Additionally, further work is needed to evaluate the utility of applying these deep learning approaches at scale on a diverse dataset. Altogether, these clinical and translational applications are increasingly motivating a need for automated methods to extract fat biomarkers from CT.

We built a fully automated abdomen and pelvis image curation and fat labeling technique using deep learning and applied it to CT scans to identify SAT and VAT. After technical validation, this technique was applied to 52 844 CT scans from 13 502 patients enrolled in the Penn Medicine Biobank (PMBB), a centralized resource of annotated blood and tissue samples linked with clinical EHR and genetic data. As additional validation of the methodology, we performed integrative analyses of the imaging traits with other phenotypic data extracted from the EHR including blood biomarkers, body mass index, and diagnoses (International Classification of Diseases–Ninth Revision [ICD-9] and International Classification of Diseases–Tenth Revision [ICD-10]) codes.

MATERIALS AND METHODS

Penn Medicine Biobank

This study used data collected from participants in the PMBB. The PMBB is a resource for advanced imaging, genetics, blood biomarkers, and other EHR data at Penn Medicine, a multihospital health system headquartered in Philadelphia, Pennsylvania. It is a detailed long-term, prospective, epidemiological study of over 50 000 volunteers containing approximately 27 485 unique diagnostic codes (ICD-9 and ICD-10) and 1 025 963 radiology studies. All patients provided informed consent to participate in the PMBB and to utilization of EHR and image data, which was approved by the Institutional Review Board of the University of Pennsylvania.

Study cohort

At the time this study was completed (January 2020), the biobank had enrolled 52 441 patients. Participants of the PMBB were included in this study if they had an abdominal and pelvis CT scan. A detailed flowchart showing the number of participants in this study, imaging studies, and number of imaging scans (multiple scans are collected per study) is shown in Figure 1. Details of the demographic and clinical summary statistics are available in Table 1.

Image analysis

A schematic of the overall approach to subcutaneous and visceral fat segmentation using deep learning is shown in Figure 2. Two deep learning neural networks were used to (1) classify 2-dimensional (2D) images of the abdomen or pelvis as showing the abdomen and then (2) segment these images for the abdominal compartment. In a final step, intensity-based thresholding criteria were applied to visceral and subcutaneous compartments to label fat in these regions. Processing was performed by utilizing distributive cloud computing

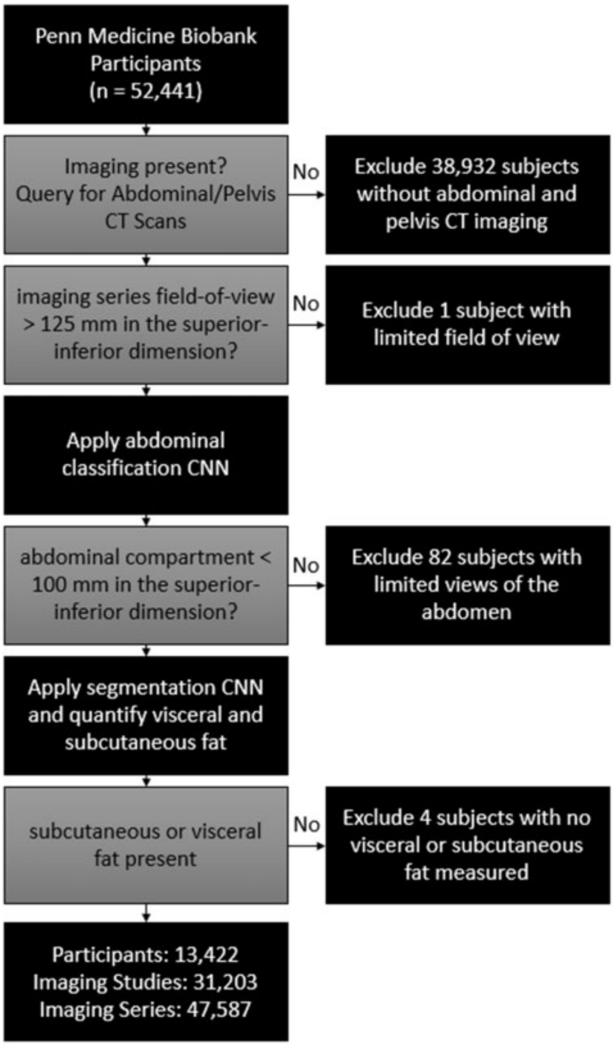
Penn Medicine Biobank Participants (n = 52,441)
Imaging present? Query for Abdominal/Pelvis CT Scans
No
Exclude 38,932 subjects without abdominal and pelvis CT imaging
imaging series field-of-view > 125 mm in the superior-inferior dimension?
No
Exclude 1 subject with limited field of view
Apply abdominal classification CNN
abdominal compartment < 100 mm in the superior-inferior dimension?
No
Exclude 82 subjects with limited views of the abdomen
Apply segmentation CNN and quantify visceral and subcutaneous fat
subcutaneous or visceral fat present
No
Exclude 4 subjects with no visceral or subcutaneous fat measured
Participants: 13,422
Imaging Studies: 31,203
Imaging Series: 47,587

Figure 1. Detailed description of the Penn Medicine Biobank participants included in this study and exclusion criteria applied at each step of analysis. An imaging study was defined as a single visit and 1 or more computed tomography (CT) scans of the abdomen or pelvis. An imaging scan was defined as a set of CT images with or without contrast; multiple scans were performed per imaging study. Abdominal classification convolutional neural network (CNN) is the network used to identify axial CT views that show the abdomen. Segmentation CNN is the network that delineates the abdominal contour. After applying exclusion criteria, association studies were performed using body mass index, blood biomarkers, and diagnostic codes.

across 50 virtual instances each equipped with an NVIDIA K80 GPU (NVIDIA, Santa Clara, CA). Please see the Supplementary Appendix for further details on our data, model training, and prediction framework.

Classification model: Identification of images showing abdominal anatomy

The first network labeled 2D slices as belonging to the abdominal cavity. Four candidate deep learning architectures were evaluated for this task, selected based on excellent performance demonstrated in the literature for classification tasks: VGG-16,[36] ResNet-50,[37] Inception V3,[38] and DenseNet-121.[39] These architectures were trained once with randomly initialized weights and then again using pretrained weights from the ImageNet dataset for a total of 8 model variants. The plus sign (eg. VGG-16+) will be used to indicate an architecture trained using ImageNet weights. The networks were trained to output a probability between 0 and 1 indicating the likelihood that the slice is within the boundary of the abdomen. To extract the boundaries from this array of probabilities we first subtracted 0.50 from every value such that such that values were in the range [-0.5, 0.5]. Next, we applied Kadane's algorithm to find the contiguous sublist of these values that gives the maximum sum.[39]

Table 1. Population characteristics for cohort (N = 13 405)

Demographics	
Age, y	57.9 ± 15.2
Sex	
Male	6,745 (50.3)
Female	6653 (49.7)
Ancestry	
European	8400 (62.7)
African	3490 (26.0)
Asian	213 (1.6)
Other/unknown	1302 (9.7)
Clinical metrics	
Height, m	1.70 ± 0.10
Weight, kg	85.6 ± 22.1
Body mass index, kg/m^2	28.8 ± 7.0
Systolic blood pressure, mm Hg	126.5 ± 12.0
Diastolic blood pressure, mm Hg	74.8 ± 7.3
Diagnoses	
Hypertension	
Yes	7103 (63.3)
No	5409 (36.7)
Diabetes	
Yes	3650 (35.0)
No	8862 (65.0)
Heart failure	
Yes	2497 (22.6)
No	10 015 (77.4)
Ischemic heart disease	
Yes	3109 (27.1)
No	9403 (72.9)
Renal failure	
Yes	3672 (35.5)
No	8840 (64.5)
Lab values	
Cholesterol, mg/dL	176.0 ± 39.1
HDL cholesterol, mg/dL	50.8 ± 15.5
LDL cholesterol, mg/dL	98.6 ± 32.2
Triglycerides, mg/dL	124.8 ± 64.2
HbA1c, %	6.4 ± 2.5
BUN, mg/dL	17.9 ± 9.8
Creatinine, mg/dL	1.01 ± 0.42

Values are mean ± SD or n (%).

BUN: blood urea nitrogen; HbA1c: glycated hemoglobin; HDL: high-density lipoprotein; LDL: low-density lipoprotein.

Training was conducted on a set of 468 scans, of which 375 scans (35 305 slices) were used for the training set and 93 scans (8775 slices) were used for the validation set. Performance of all 8 networks (4 distinct architectures, trained with and without pretrained weights) was then evaluated for its sensitivity, specificity, and accuracy on a separate testing set of 100 scans, which were randomly selected from the PMBB. For each network, these metrics were first calculated individually on each scan, and then the metrics were averaged across all 100 scans. For both the classification and

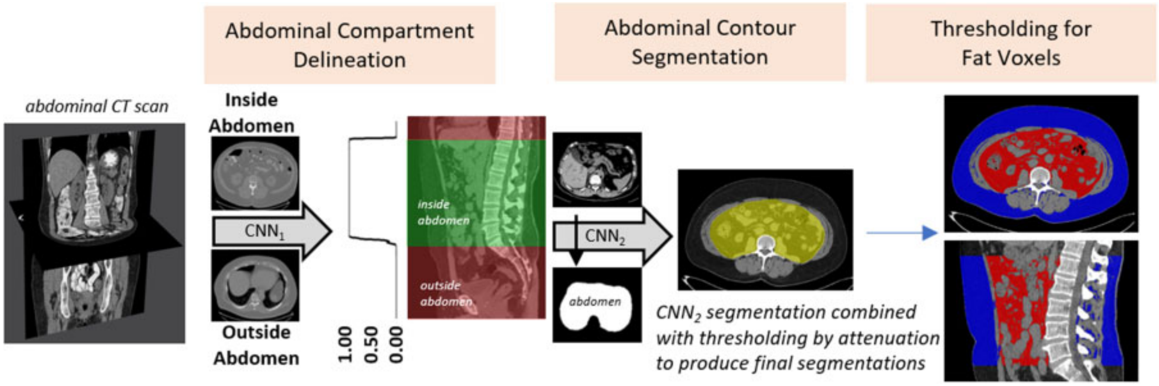
Abdominal Compartment Delineation
Abdominal Contour Segmentation
Thresholding for Fat Voxels
abdominal CT scan
Inside Abdomen
CNN_1
Outside Abdomen
1.00
0.50
0.00
inside abdomen
outside abdomen
CNN_2
abdomen
CNN_2 segmentation combined with thresholding by attenuation to produce final segmentations

Figure 2. Automated extraction of abdominal fat from computed tomography (CT) scans. Abdominal classification network (convolutional neural network 1 [CNN_1]) classifies 2-dimensional image slices as belonging inside the abdomen or outside the abdomen. Segmentation network (CNN_2) delineates the border of the inner abdominal contour. Fat voxels are identified based on CT attenuation and those within the contour represent visceral and those outside represent subcutaneous adipose tissue.

segmentation tasks, metrics were compared using pairwise *t* tests with Bonferroni multiple comparison correction and a level of significance of 0.05. The highest performing network was selected, and 5-fold cross-validation was then performed on this network using all available 568 scans.

Segmentation model architecture: Labeling of subcutaneous and visceral fat pixels

The second network delineated SAT and VAT from axial 2D slices. Three candidate deep learning architectures were selected based on excellent performance in the literature for segmentation tasks: U-Net,[40] Deep Lab V3 using Xception encoder,[41,42] and Deep Lab V3 using MobileNet V2 encoder.[41,43] The networks output a probability for each voxel indicating the probability that it belongs to the foreground, and probabilities $\geq$0.50 are attributed to the foreground. Additional postprocessing steps include only preserving the largest connected component and filling any holes for each slice.

Training was conducted on a set of 62 scans with 50 scans (2059 slices) randomly selected for the training set and 12 scans (498 slices) for the validation set. Performance of these networks was evaluated on a separate testing set of 20 scans, which were randomly selected from the PMBB. Region-of-interest area overlap ratios (Dice scores)[44] were calculated to measure agreement between manual and automatic segmentations for the abdominal contour as well as SAT and VAT. The highest performing network was selected, and then 5-fold cross-validation was conducted on this network using all 82 scans. For additional evaluation, 100 abdomen and pelvis CT scans were randomly selected from the PMBB, and VAT and SAT was manually segmented on a single slice between L3 and L4. These values were compared with automatically derived measurements from the highest-performing model.

Association studies

BMI correlation analysis

The highest performing networks for both the classification and segmentation tasks were then used to process all 31 419 studies. The convolutional neural network (CNN)–derived metrics for SAT and VAT were further validated by comparing these values with clinically assessed BMI values. The most recent BMI measurement was attributed to each scan, and measurements acquired >365 days from the scan were excluded. Pearson's correlation coefficient was computed to measure the degree of association. In all association analyses, we utilized the average metric area of SAT or VAT across all slices. To associate a single fat value to each patient, we took the median (or mean if exactly 2) value for studies with multiple scans or for patients with multiple studies. This approach for attributing a single image-derived phenotype to a patient was utilized in all association studies.

Phenome-wide association study

A phenome-wide association study (PheWAS) was performed to investigate the phenotypic associations of having a higher VAT-SAT ratio (VSR). ICD-10 codes were first mapped to ICD-9 codes using the 2017 general equivalency mapping. Next, ICD-9 codes were aggregated into phecodes using the PheWAS R package to create 1816 phecodes. Patients with at least 2 occurrences of a phecode are considered cases, those with none are control subjects, and those with 1 are treated as missing. Phecodes with <100 cases were excluded. Logistic regression was then performed with each phecode as the outcome and VSR as a predictor. Regression was performed controlling for the covariates of age, sex, and race. Bonferroni multiple comparison correction was used to determine the level of significance.

Relationship between lab values and VSR

To investigate the relationship between VSR and common clinical laboratory studies, VSR values were organized into quartiles. Because VSR is associated with both sex and race, patients were first stratified into 4 categories based on sex and race. Patients with the highest and lowest 25% of values within each of these categories were identified and the distribution of laboratory measures between the high and low groups was compared. Lab values acquired >90 days and BMI measurements acquired >365 days from the scan date were excluded. To compare the distributions, a Wilcox rank sum test was performed.

RESULTS

Patient cohort acquisition and analysis

After applying exclusion criteria, 31 419 studies containing a total of 52 844 scans corresponding to 13 502 patients were analyzed to identify SAT and VAT. Processing was performed in 5 hours and 8 minutes (0.35 s/scan) by using parallel processing. The total runtime across all 50 instances was 201 hours and 43 minutes

Table 2. Performance metrics for classifier networks

Type	Inferior	Superior	Sensitivity	Specificity	Accuracy
VGG-16	0.40 ± 0.65	0.55 ± 1.40	0.988 ± 0.031	0.991 ± 0.015	0.990 ± 0.014
DenseNet-121	0.61 ± 1.05	0.67 ± 1.32	0.978 ± 0.042	0.990 ± 0.012	0.986 ± 0.017
ResNet-50	0.82 ± 1.11	1.03 ± 1.76	0.972 ± 0.053	0.982 ± 0.024	0.978 ± 0.022
Inception V3	0.42 ± 0.64	0.38 ± 1.09	0.986 ± 0.034	0.987 ± 0.017	0.986 ± 0.016
VGG-16+	0.42 ± 0.64	0.38 ± 1.09	0.989 ± 0.025	0.993 ± 0.014	0.991 ± 0.013
DenseNet-121+	0.49 ± 1.24	0.30 ± 1.03	0.987 ± 0.032	0.995 ± 0.010	0.992 ± 0.014
ResNet-50+	0.36 ± 0.73	0.41 ± 1.02	0.988 ± 0.030	0.993 ± 0.013	0.991 ± 0.013
Inception V3+	0.42 ± 0.56	0.54 ± 1.37	0.985 ± 0.035	0.993 ± 0.011	0.990 ± 0.015
VGG-16 + 5-Fold	0.32 ± 0.88	0.41 ± 0.98	0.988 ± 0.029	0.994 ± 0.013	0.992 ± 0.014

Values are mean ± SD.

Table 3. Performance metrics for segmentation networks

	Abdominal Contour	Subcutaneous Fat	Visceral Fat
U-Net	0.980 ± 0.008	0.998 ± 0.003	0.991 ± 0.007
DeepLab+MobileNet V2	0.975 ± 0.008	0.998 ± 0.002	0.986 ± 0.009
DeepLab+Xception	0.979 ± 0.006	0.998 ± 0.001	0.990 ± 0.006
U-Net 5-fold	0.972 ± 0.052	0.978 ± 0.072	0.997 ± 0.008
U-Net Extended Training	0.982 ± 0.007	0.998 ± 0.002	0.992 ± 0.006

Values are mean ± SD.

(13.7 s/scan). Following the exclusion of additional scans based on the size of abdominal window and volume of fat detected, 31 158 studies containing 47 470 scans corresponding to 13 405 patients were utilized in association studies (Figure 1).

Image analysis: Technical validation

Localizer network

Without the use of pretrained weights, the VGG-16, ResNet-50, Inception V3, and DenseNet-121 networks converged after 27, 59, 54, and 40 epochs, respectively. When using ImageNet weights, the networks converged more quickly after 9, 17, 8, and 17 epochs, respectively. The performance of the classifier networks is shown in Table 2. All architectures achieved a sensitivity >0.97, specificity >0.98, and accuracy >0.98. Using pretrained weights for model training did not significantly increase sensitivity for any of the architectures ($P \geq .081$), but specificity ($P = 1.4 \times 10^{-5}$) and accuracy ($P = 3.9 \times 10^{-7}$) did increase for ResNet-50. When comparing the metrics between all architectures, VGG-16+ had a significantly greater average sensitivity ($P = .025$), specificity ($P = 5.9 \times 10^{-5}$), and accuracy ($P = 2.5 \times 10^{-7}$) than ResNet-50, but pairwise *t* tests between VGG-16+ and the other architectures demonstrated no difference ($P \geq .18$). Regarding runtime, the VGG-16, ResNet-50, Inception V3, and DenseNet-121 models took on average 4.2, 3.6, 5.0, and 4.3 seconds, respectively, per scan to process the 100 testing cases.

VGG-16+ was selected as the architecture of choice based on its noninferior performance, simplicity of design, and fast runtime. The automated method was on average within one 5-mm slice from the manually selected slice for both the superior (0.4 ± 1.1 slice) and inferior (0.4 ± 0.6 slice) borders. For the superior border, the prediction ranged from 3 slices below to 3 slices above the manual label. For the inferior border, they ranged from 6 slices below to 8 slices above (1 vertebral level) the manual label. Before application of the maximum sub-list algorithm, the VGG-16+ algorithm had a sensitivity of 0.99 ± 0.03, specificity of 0.99 ± 0.01, and accuracy of 0.99 ± 0.01. These metrics were without significant change after application of Kadane's algorithm—*P* values of .82, .65, and .92, respectively. Fivefold cross-validation on the VGG-16+ architecture was then performed using all available 568 studies. It demonstrated excellent sensitivity, specificity, and accuracy values with average values of 0.99 for all 3 metrics (Table 2). These metrics obtained during cross-validation were not significantly different from those obtained with VGG-16+ on the testing set ($P \geq .32$).

Segmentation network

The U-Net, DeepLab+MobileNet V2, and DeepLab+Xception converged after 39, 70, and 63 epochs, respectively. When assessing performance on the testing set of 20 scans, all architectures achieved mean Dice values ≥0.98 for the abdominal contour as well as the SAT and VAT regions (Table 3). When conducting pairwise comparison between the 3 algorithms, there was no significant difference in the means for any of the metrics ($P \geq .16$). Regarding runtime, U-Net, DeepLab+MobileNet V2 and DeepLab+Xception ran in 7.3, 7.3, and 8.5 seconds per case, respectively. The U-Net architecture was selected for its non-inferior performance, runtime efficiency, and simplicity of design. 5-fold cross-validation was then performed for the U-Net architecture, and it obtained mean Dice values of at least 0.97 for all 3 metrics (Table 3). The metrics obtained during cross-validation were also not significantly different from those obtained with U-Net on the original testing set ($P \geq .11$).

In our experience, while extending the duration of training does not yield a significant improvement in Dice metrics, it can result in better performance on edge cases, provided that the model is not allowed to overfit. For this reason, we trained the U-Net for 345 epochs and selected epoch 326 based on Dice values for the validation set; these weights were used for processing studies at scale. Dice metrics for evaluation of this model are shown in Table 3 and representative segmentation results are shown in Figure 3. Scatterplots and Bland-Altman plots showing evaluation results for the single

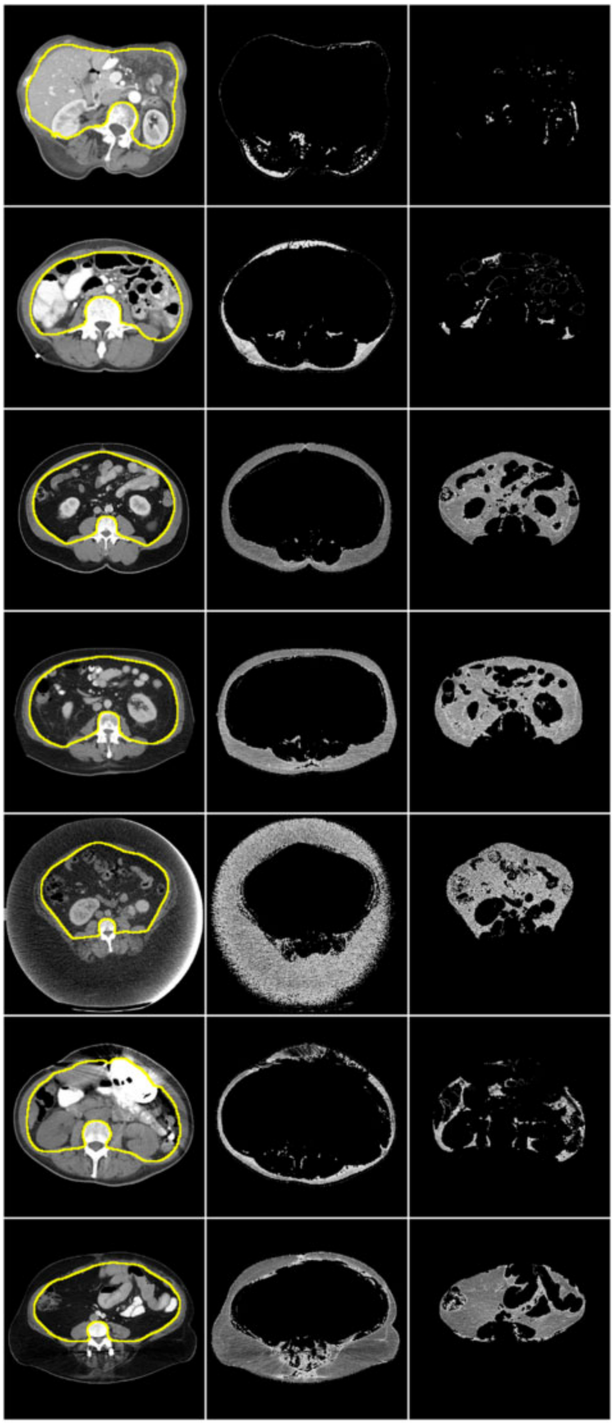

Figure 3. Representative segmentation results from 6 different patients. First column shows computed tomography slice with yellow line indicating the convolutional neural network–predicted contour. From this contour, our algorithm identifies subcutaneous fat (second column) and visceral fat (third column). Row 5 shows a patient with beam hardening artifact from a left ventricular assist device, and row 6 shows a patient with subcutaneous scar tissue from a spinal fusion surgery.

slice segmentations on 100 different patients is shown in Figure 4A to 4D. Intraclass correlation coefficients of 0.9999 ($P < 2 \times 10^{-16}$) and 0.9998 ($P < 2 \times 10^{-16}$) were achieved for the prediction of SAT and VAT, respectively. There was a significant bias of 1.4 ± 1.6 cm^2 ($P = 3.1 \times 10^{-15}$) for SAT and -1.4 ± 1.6 cm^2 ($P = 3.1 \times 10^{-15}$) for VAT. For this analysis, the average areas were 298.7 ± 167.1 cm^2 for SAT and 153.5 ± 109.6 cm^2 for VAT.

Association studies

Association studies were conducted investigating the relationship between the CNN-derived fat values and BMI, clinical lab values, and billing codes. The relationships between the CNN-derived fat values and BMI are shown in Figure 4E and 4F. There was a significant correlation between BMI and both subcutaneous (r = 0.876; $P < 2 \times 10^{-16}$) and visceral (r = 0.522; $P < 2 \times 10^{-16}$) fat.

Next, we compared the distribution of lab values and BMI measurements for patients in the bottom quartile of VSR values with those in the top quartile. For the high-VSR group there was a significant increase in triglycerides ($P = 5.9 \times 10^{-10}$), glycated hemoglobin ($P = 2.0 \times 10^{-4}$), blood urea nitrogen ($P = 1.0 \times 10^{-14}$), creatinine ($P = 9.8 \times 10^{-55}$), and BMI ($P = 1.7 \times 10^{-5}$). There was also a significant decrease in high-density lipoprotein (HDL) ($P = 0.0014$). Density plots showing the distribution of values for the 2 groups are shown in Figure 5.

A PheWAS of VSR revealed significant associations with several pathologies. The plot is shown in Figure 6. The strongest association was with diabetes mellitus ($P = 1.7 \times 10^{-23}$). There were multiple hits for other endocrine disorders related to diabetes or lipid metabolism. Within the circulatory system, the strongest signals were for hypertension ($P = 2.5 \times 10^{-19}$) and hypertension-related complications. In the renal system, there were 7 significant hits, including chronic kidney disease ($P = 5.4 \times 10^{-18}$), renal failure ($P = 1.7 \times 10^{-15}$), and renal transplant ($P = 1.6 \times 10^{-12}$).

DISCUSSION

In this article, we present a fully automated approach to accurately quantify abdominal fat from clinical CT scans. We provide a rigorous evaluation of several prominent deep learning architectures for this task as well as an evaluation on the use of transfer learning by using ImageNet weights. Following architecture selection, we utilized distributive processing in a cloud computing environment to quantify full abdominal volumes of visceral and subcutaneous fat from 52,844 scans in 5 hours (~172 scans/min). This demonstrates the efficiency of a fully trained deep learning network to label abdominal CT images at a high rate which is necessary to perform large-scale research applications or to support automatic, quantitative reporting of abdominal fat by radiologists. To our knowledge, this is the first automatic technique for 3-dimensional CT abdominal fat segmentation to be associated at scale to unbiased EHR data from an academic biobank.

During the technical validation of our methods, the results showed outstanding agreement between manual and automated measurements of fat. There was a significant bias of 1.4 cm^2 for both SAT and VAT. However, compared to the average areas of 299 cm^2 for SAT and 154 cm^2 for VAT in our testing set, this bias represented a small error (<1%). While previous studies that utilized radial projections or shape analysis were not tested on patients with extreme body habitus,[16,21] our method demonstrated excellent performance on both the randomly selected testing set and during 5-fold validation, which included testing on many studies that were intentionally selected for either noise or extreme body habitus. Figure 3 highlights the fault tolerance of the algorithm to variations in patient body habitus and BMI. Indeed, when representative diverse cases

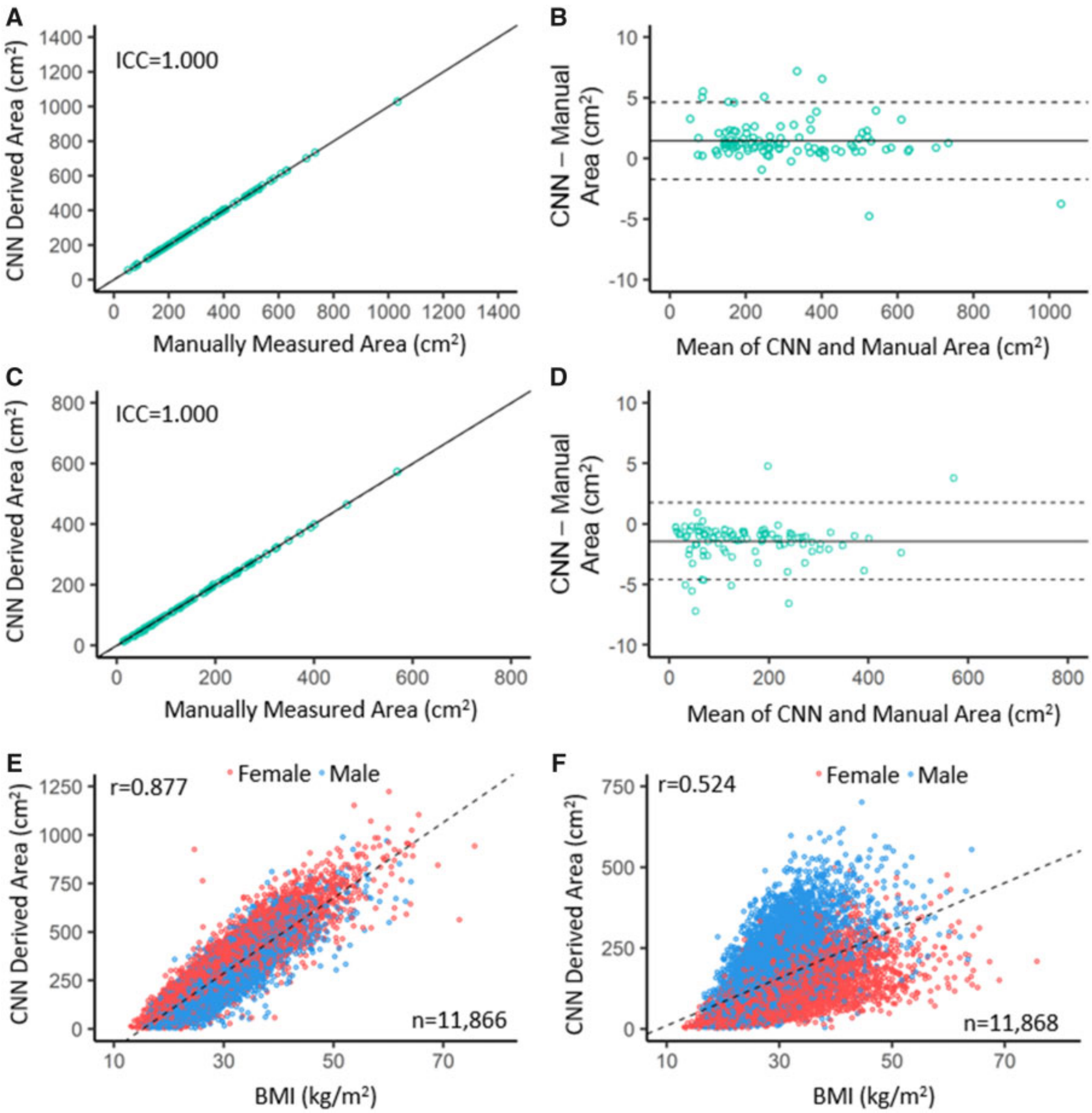

Figure 4. Scatterplots and Bland-Altman plots comparing convolutional neural network (CNN) to manually derived area for (A, B) subcutaneous and (C, D) visceral fat, in which fat was derived from a single slice on 100 randomly selected scans. (E, F) Scatterplots of CNN-derived area vs body mass index for (E) subcutaneous and (F) visceral fat, in which fat was derived from all available biobank scans. ICC: intraclass correlation coefficient.

are included in the training data, data-driven segmentation correctly quantified fat area when there was no visible SAT or when body habitus or postsurgical changes caused significant artifacts. Work by Weston et al[33] applied a similar U-Net CNN for body composition analysis to manual selected CT slices and achieved Dice scores of 0.98 ± 0.03 for SAT and 0.94 ± 0.12 for VAT. These are comparable to our values of 0.998 ± 0.002 for SAT and 0.992 ± 0.006 for VAT. Given that we used a similar method, the small improvement in our approach is likely due to the diversity of pathology in our training data.

We found that the association between SAT or VAT and BMI was strong and similar compared with previous studies that found associations of r = 0.73 to 0.93 for SAT and r = 0.61 to 0.77 for VAT.[7,45,46] In agreement with these studies, we found a stronger association between BMI and SAT than between BMI and VAT. When comparing lab values between patients with high and low values of VSR, the high-VSR group had significantly higher triglycerides and lower HDL but no significant change in low-density lipoprotein. This was consistent with previous findings that VSR was positively associated with triglycerides, negatively associated with HDL, and showed no significant association with low-density lipoprotein.[19] Glycated hemoglobin levels were significantly elevated in our analysis, which aligns with a known association with diabetes.[4] While an association between VSR and liver enzymes has been reported,[47] our analysis did not find significant associations for AST or ALT. Our finding of increased creatinine and blood urea nitrogen in the higher-VSR group is consistent with previous findings that increased VAT is associated with decreased renal function and progression to end-stage renal disease, particularly in diabetic kidney disease.[48]

By investigating association between VSR and phecodes, we were able to interrogate disease associations in an unbiased manner.

As expected, we found numerous associations with diabetic, hypertensive, and kidney disease pathologies. There were some unexpected associations with transplants and human immunodeficiency virus infections. The association with human immunodeficiency virus is likely due to lipodystrophy, which is commonly seen in these patients.[49] Similarly, the association with transplants is likely secondary to the use of corticosteroids following transplant, which is known at high doses to increase the amount of VAT.[50] The negative association with bariatric surgery is likely because patients who get bariatric surgery have significant depots of subcutaneous fat.

While the VGG-16 and U-Net networks are frequently applied for medical imaging applications, there are multiple technical and translational advances demonstrated in this article. These major contributions include (1) developing a 2-step classification-segmentation pipeline that efficiently processes scans without the need for any manual input, (2) providing rigorous comparisons of multiple deep learning architectures for this application, and (3) conducting association studies between image features and phenotypes in an academic biobank to provide both additional validation as well as to highlight the utility of applying our method at scale.

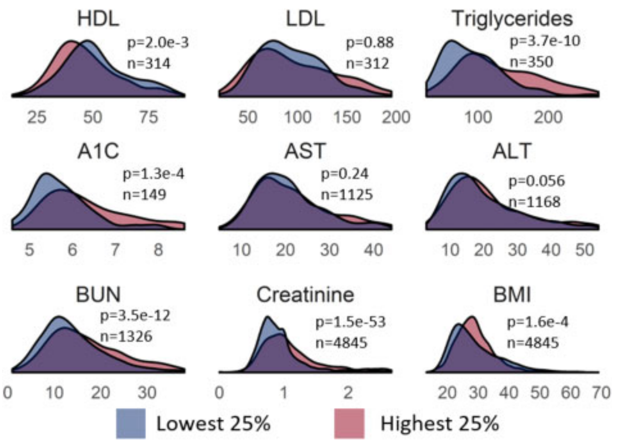

Figure 5. Biomarker distributions dichotomized by visceral adipose tissue-to-subcutaneous adipose tissue ratio values. A1C: glycated hemoglobin; ALT: alanine aminotransferase; AST: aspartate aminotransferase; BMI: body mass index; BUN: blood urea nitrogen; HDL: high-density lipoprotein; LDL: low-density lipoprotein.

There are several limitations to this study. While the attenuation range for fat has been defined in literature, CT scans acquired over several decades may contain artifacts or utilize reconstruction algorithms that distort attenuation. Specifically, implanted devices can distort attenuation by creating beam hardening artifacts or scar tissue can be introduced, changing the tissue attenuation by physiologic means. Given our automated approach, it is also possible that scans do not reach the inferior or superior borders of the abdomen, and this could skew the resulting fat values. This study also derived disease phenotypes from EHR billing codes, which are often incomplete. As this was a retrospective cohort, there could be significant selection bias for sicker patients or certain diseases based on biobank recruitment methods. Further work should be performed to investigate the associated phenotypes to refine our understanding and identify any causative relationships. Additionally, there is great potential in the utilization of this method in the context of a biobank such as the PMBB in which genetic information is available. This may provide greater insight into the mechanism of pathogenicity for VAT, which would be of great interest to the scientific community.

In conclusion, this study presents a fully automated method for the quantification of abdominal fat that functions with high accuracy and can be applied efficiently in a cloud computing environment. This method has been validated through traditional technical approaches and by integration of our results with clinical data. This integration further highlights how autonomous image trait

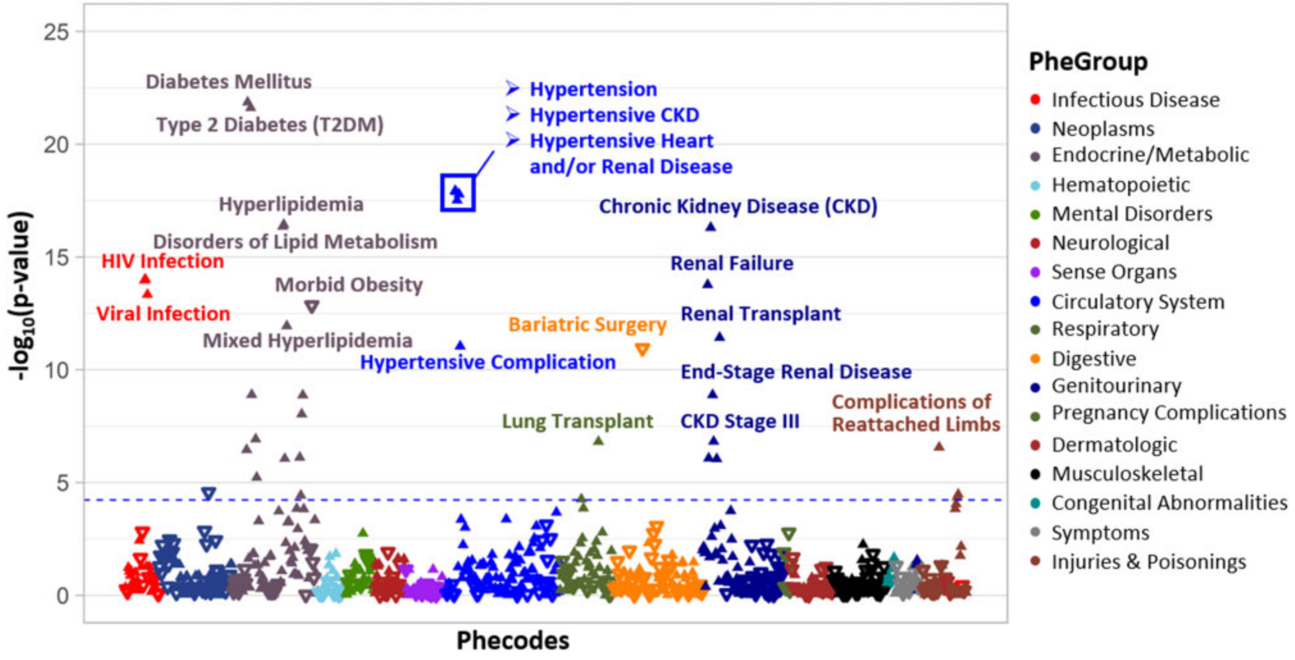

Figure 6. Phenome-wide association study of the visceral adipose tissue-to-subcutaneous adipose tissue ratio. The blue line indicates the level of significance with Bonferroni multiple-comparison correction. Upward-facing triangles indicate a positive association and downward-facing triangles indicate a negative association. CKD: chronic kidney disease; HIV: human immunodeficiency virus; T2DM: type 2 diabetes mellitus.

quantification can facilitate translational research especially in the context of an academic biobank.

FUNDING

This work was supported by the Sarnoff Cardiovascular Research Foundation (to MM); National Institutes of Health National Center for Advancing Translational Studies UL1TR001878; National Institutes of Health/National Heart, Lung, and Blood Institute R01 HL137984, R01 AA026302-02, P30 DK0503060 (to RC); and the Penn Center for Precision Medicine.

AUTHOR CONTRIBUTIONS

WRW, DR, RC, and MM obtained the funding. WRW, DR, RC, and MM were responsible for the concept and design of the study. WRW, DR, MM, MS, DM, and AB were involved in patient identification and data procurement from the clinical workflow. MM, QJ, JC, DT, and MR were involved in the process of annotating ground-truth data. WRW, MM, HS and HL were involved in model training and evaluation. WRW, MM, HS, MV, YK, and SD were involved in the statistical analysis. WRW, DR and MM drafted the manuscript, and all authors revised and approved the final manuscript.

COMPETING INTERESTS STATEMENT

All authors have no competing interests to declare.

DATA AVAILABILITY STATEMENT

Data is available upon reasonable request to investigators.

REFERENCES

1. Frezza EE, Wachtel MS, Chiriva-Internati M. Influence of obesity on the risk of developing colon cancer. *Gut* 2006; 55 (2): 285–91.
2. Van Gaal LF, Mertens IL, De Block CE. Mechanisms linking obesity with cardiovascular disease. *Nature* 2006; 444 (7121): 875–80.
3. Burkhauser RV, Cawley J. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. *J Health Econ* 2008; 27 (2): 519–29.
4. Bergman RN, Kim SP, Catalano KJ, *et al.* Why visceral fat is bad: mechanisms of the metabolic syndrome. *Obesity (Silver Spring)* 2006; 14 (2S): 16S–9S.
5. Funahashi T, Nakamura T, Shimomura I, *et al.* Role of adipocytokines on the pathogenesis of atherosclerosis in visceral obesity. *Intern Med* 1999; 38 (2): 202–6.
6. Matsuzawa Y, Funahashi T, Nakamura T. The concept of metabolic syndrome: contribution of visceral fat accumulation and its molecular mechanism. *J Atheroscler Thromb* 2011; 18 (8): 629–39.
7. Janssen I, Heymsfield SB, Allison DB, *et al.* Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *Am J Clin Nutr* 2002; 75 (4): 683–8.
8. Janssen I, Katzmarzyk PT, Ross R. Body mass index, waist circumference, and health risk: evidence in support of current National Institutes of Health guidelines. *Arch Intern Med* 2002; 162 (18): 2074–9.
9. Pouliot M-C, Després J-P, Lemieux S, *et al.* Waist circumference and abdominal sagittal diameter: best simple anthropometric indexes of abdominal visceral adipose tissue accumulation and related cardiovascular risk in men and women. *Am J Cardiol* 1994; 73 (7): 460–8.
10. Lofgren I, Herron K, Zern T, *et al.* Waist circumference is a better predictor than body mass index of coronary heart disease risk in overweight premenopausal women. *J Nutr* 2004; 134 (5): 1071–6.
11. Staiano AE, Reeder BA, Elliott S, *et al.* Body mass index versus waist circumference as predictors of mortality in Canadian adults. *Int J Obes* 2012; 36 (11): 1450–4.
12. Bosy-Westphal A, Booke C-A, Blöcker T, *et al.* Measurement site for waist circumference affects its accuracy as an index of visceral and abdominal subcutaneous fat in a Caucasian population. *J Nutr* 2010; 140 (5): 954–61.
13. Kinkorová J. Biobanks in the era of personalized medicine: objectives, challenges, and innovation. *EPMA J* 2015; 7 (1): 4.
14. Jiang H, Nachum O. Identifying and correcting label bias in machine learning. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), Volume 108*; 2020.
15. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *arXiv Preprint arXiv*, doi: http://arxiv.org/abs/1908.09635, 17 Sep 2019, preprint: not peer reviewed.
16. Zhao B, Colville J, Kalaigian J, *et al.* Automated quantification of body fat distribution on volumetric computed tomography. *J Comput Assist Tomogr* 2006; 30 (5): 777–83.
17. Pednekar A, Bandekar AN, Kakadiaris IA, Naghavi M. Automatic segmentation of abdominal fat from CT data. In: *Proceedigns of the 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*; IEEE; 2005; New York, NY.
18. Kim YJ, Lee SH, Kim TY, *et al.* Body fat assessment method using CT images with separation mask algorithm. *J Digit Imaging* 2013; 26 (2): 155–62.
19. Cha EDK, Veturi Y, Agarwal C, *et al.* Using adipose measures from health care provider-based imaging data for discovery. *J Obes* 2018; 2018: 3253096.
20. Positano V, Gastaldelli A, Sironi A. M, *et al.* An accurate and robust method for unsupervised assessment of abdominal fat by MRI. *J Magn Reson Imaging* 2004; 20 (4): 684–9.
21. Chung H, Cobzas D, Birdsell L, Lieffers J, Baracos V. Automated segmentation of muscle and adipose tissue on CT images for human body composition analysis. In: *Proceedings of Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling, Volume 7261*; SPIE; March 13, 2009; Bellingham, WA.
22. Sadananthan SA, Prakash B, Leow MK-S, *et al.* Automated segmentation of visceral and subcutaneous (deep and superficial) adipose tissues in normal and overweight men. *J Magn Reson Imaging* 2015; 41 (4): 924–34.
23. Popuri K, Cobzas D, Esfandiari N, *et al.* Body composition assessment in axial CT images using FEM-based automatic segmentation of skeletal muscle. *IEEE Trans Med Imaging* 2016; 35 (2): 512–20.
24. Agarwal C, Dallal AH, Arbabshirani MR, Patel A, Moore G. Unsupervised quantification of abdominal fat from CT images using Greedy Snakes. In: *Proceedings of Medical Imaging 2017: Image Processing, Volume 10133*; SPIE; February 24, 2017; Bellingham, WA.
25. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. *Int J Comput Vision* 1988; 1 (4): 321–31.
26. Mishra AK, Fieguth PW, Clausi DA. Decoupled active contour (DAC) for boundary detection. *IEEE Trans Pattern Anal Mach Intell* 2011; 33 (2): 310–24.
27. Sharma N, Ray AK, Shukla KK, *et al.* Automated medical image segmentation techniques. *J Med Phys* 2010; 35 (1): 3.
28. Jiang F, Li H, Hou X, *et al.* Abdominal adipose tissues extraction using multi-scale deep neural network. *Neurocomputing* 2017; 229: 23–33.
29. Christ PF, Elshaer MEA, Ettlinger F, *et al.* Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: *MICCAI 2016: Medical Image Computing and Computer-Assisted Intervention*; 2016: 415–23.
30. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542 (7639): 115–8.
31. Gulshan V, Peng L, Coram M, *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316 (22): 2402–10.
32. Ehteshami Bejnordi B, Veta M, Johannes van Diest P; *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; 318 (22): 2199–210.

33. Weston AD, Korfiatis P, Kline TL, *et al.* Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology* 2019; 290 (3): 669–79.
34. Park HJ, Shin Y, Park J, *et al.* Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. *Korean J Radiol* 2020; 21 (1): 88–100.
35. Wang Y, Qiu Y, Thai T, *et al.* A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images. *Comput Methods Programs Biomed* 2017; 144: 97–104.
36. Hassan M. *VGG16: Convolutional Network for Classification and Detection*. 2019. https://neurohive.io/en/popular-networks/vgg16/#:∼:text=VGG16%20is%20a%20convolutional%20neural%20network%20model%20proposed%20by%20K.&text=Zisserman%20from%20the%20University%20of,images%20belonging%20to%201000%20classes. Accessed February 2, 2021.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016.
38. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *arXiv*, doi: http://arxiv.org/abs/1512.00567, 11 Dec 2015, preprint: not peer reviewed.
39. Takaoka T. Efficient algorithms for the maximum subarray problem by distance matrix multiplication. *Electr Notes Theor Compu Sci* 2002; 61: 191–200.
40. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: *MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention*; 2015: 234–41.
41. Chen L-C, Papandreou G, Kokkinos I, *et al.* DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 2018; 40 (4): 834–48.
42. Chollet F. Xception: Deep learning with depthwise separable convolutions, 2016. *arXiv*, doi: http://arxiv.org/abs/1610.02357, 4 Apr 2017, preprint: not peer reviewed.
43. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018: 4510–20.
44. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; 26 (3): 297–302.
45. Kim JH, Doo SW, Cho KS, *et al.* Which anthropometric measurements including visceral fat, subcutaneous fat, body mass index, and waist circumference could predict the urinary stone composition most? *BMC Urol* 2015; 15 (1): 17.
46. Camhi SM, Bray GA, Bouchard C, *et al.* The relationship of waist circumference and BMI to visceral, subcutaneous, and total body fat: sex and race differences. *Obesity* 2011; 19 (2): 402–8.
47. Jung C-H, Rhee E-J, Kwon H, *et al.* Visceral-to-subcutaneous abdominal fat ratio is associated with nonalcoholic fatty liver disease and liver fibrosis. *Endocrinol Metab* 2020; 35 (1): 165–76.
48. Kim SR, Yoo JH, Song HC, *et al.* Relationship of visceral and subcutaneous adiposity with renal function in people with type 2 diabetes mellitus. *Nephrol Dial Transplant* 2011; 26 (11): 3550–5.
49. Carr A. HIV lipodystrophy: risk factors, pathogenesis, diagnosis and management. *AIDS* 2003; 17: S141–8.
50. Rockall AG, Sohaib SA, Evans D, *et al.* Computed tomography assessment of fat distribution in male and female patients with Cushing's syndrome. *Eur J Endocrinol* 2003; 149 (6): 543–68.