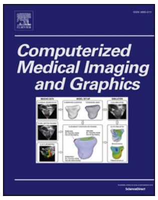
Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



Uncertainty-aware body composition analysis with deep regression ensembles on UK Biobank MRI

Taro Langner [a,*], Fredrik K. Gustafsson [b], Benny Avelin [c], Robin Strand [b], Håkan Ahlström [a,d], Joel Kullberg [a,d]

[a] Uppsala University, Department of Surgical Sciences, Uppsala, Sweden
[b] Uppsala University, Department of Information Technology, Uppsala, Sweden
[c] Uppsala University, Department of Mathematics, Uppsala, Sweden
[d] Antaros Medical AB, BioVenture Hub, Mölndal, Sweden

ARTICLE INFO

Keywords:
MRI
UK Biobank
Neural networks
Biomarkers

ABSTRACT

Along with rich health-related metadata, medical images have been acquired for over 40,000 male and female UK Biobank participants, aged 44–82, since 2014. Phenotypes derived from these images, such as measurements of body composition from MRI, can reveal new links between genetics, cardiovascular disease, and metabolic conditions. In this work, six measurements of body composition and adipose tissues were automatically estimated by image-based, deep regression with ResNet50 neural networks from neck-to-knee body MRI. Despite the potential for high speed and accuracy, these networks produce no output segmentations that could indicate the reliability of individual measurements. The presented experiments therefore examine uncertainty quantification with mean-variance regression and ensembling to estimate individual measurement errors and thereby identify potential outliers, anomalies, and other failure cases automatically. In 10-fold cross-validation on data of about 8500 subjects, mean-variance regression and ensembling showed complementary benefits, reducing the mean absolute error across all predictions by 12%. Both improved the calibration of uncertainties and their ability to identify high prediction errors. With intra-class correlation coefficients (ICC) above 0.97, all targets except the liver fat content yielded relative measurement errors below 5%. Testing on another 1000 subjects showed consistent performance, and the method was finally deployed for inference to 30,000 subjects with missing reference values. The results indicate that deep regression ensembles could ultimately provide automated, uncertainty-aware measurements of body composition for more than 120,000 UK Biobank neck-to-knee body MRI that are to be acquired within the coming years.

1. Introduction

UK Biobank studies more than half a million volunteers by collecting data on blood biochemistry, genetics, questionnaires on lifestyle, and medical records (Sudlow et al., 2015).

For 100,000 participants, the ongoing examinations also include medical imaging, such as dedicated MRI of the brain, heart, liver, pancreas, and the entire body from neck to knee (Littlejohns et al., 2020). Ongoing repeat imaging for 70,000 subjects will furthermore enable longitudinal studies over two or more years. Image-derived phenotypes, such as measurements of body composition and organ volumes, hold the potential for non-invasive studies of aging, cardiovascular, and metabolic conditions at large scale within this cohort.

The relationship between obesity, type-2 diabetes, and nonalcoholic fatty liver disease is of particular interest due to their high prevalence and associated adverse health effects (Wilman et al., 2017; Linge et al., 2018). Depending on genetic and environmental factors, body fat can accumulate in organs, abdominal depots, and muscle infiltrations, all of which have specific effects on health outcomes. Ongoing work is therefore concerned with acquiring measurements of liver fat content (Wilman et al., 2017), muscle volumes, and adipose tissue depots (West et al., 2016; Linge et al., 2018) with manual and semi-automated techniques (Borga, 2018). Recent works also proposed fully-automated techniques with neural networks for segmentation, which have been applied to the heart (Bai et al., 2018), kidney (Langner et al., 2020a), pancreas (Basty et al., 2020; Bagur et al., 2020), and liver (Irving et al.,

[*] Corresponding author.
E-mail address: taro.langner@surgsci.uu.se (T. Langner).

https://doi.org/10.1016/j.compmedimag.2021.101994
Received 16 April 2021; Received in revised form 6 September 2021; Accepted 7 September 2021
Available online 23 September 2021
0895-6111/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2017), but also the iliopsoas muscles (Fitzpatrick et al., 2020), spleen, adipose tissues, and more (Liu et al., 2021). Similar to the latter, neural networks have also been proposed for segmentation of adipose tissues in other studies involving computed tomography (CT) (Wang et al., 2017; Weston et al., 2019) and MRI (Langner et al., 2019; Estrada et al., 2020; Küstner et al., 2020).

Apart from semantic segmentation, neural networks can also be trained for image-based regression, predicting numerical measurement values without any need for explicit delineations. In medical imaging, deep regression has gained attention for analyses of human age in MRI of the brain (Cole et al., 2018), volume measurements of the heart (Xue et al., 2017), and blood pressure, sex, and age in retinal fundus photographs (Poplin et al., 2018). On UK Biobank neck-to-knee body MRI, deep regression can quantify human age and liver fat, but also various measurements of body composition. For the latter, its accuracy can exceed the agreement between established gold standard techniques (Langner et al., 2020b).

This type of deep regression requires no ground truth segmentations and can measure abstract properties by training on numerical reference values from arbitrary sources. However, the lack of output segmentations poses a limitation, as the predicted numerical values give no indication of confidence or reliability. Previous work examined the underlying relevant image features with saliency analysis, but only provided interpretations on cohort level without attempting to estimate individual measurement errors.

Recent advances in the field of uncertainty quantification have the potential to address some of these concerns by providing an error estimate for each individual measurement (Ghahramani, 2015). High uncertainty could accordingly alert researchers or clinical operators to anomalies, outliers, or other failure cases of these systems (Kendall and Gal, 2017). Among various proposed methods, such as Bayesian inference with Markov chain Monte-Carlo techniques (Neal, 2012) and more computationally viable approximations that apply dropout at test time (Gal and Ghahramani, 2016), recent work reported superior behavior for deep ensembling strategies (Gustafsson et al., 2020; Ovadia et al., 2019; Ashukha et al., 2020). These approaches provide *predictive uncertainty* by training multiple neural networks to each predict not only a point estimate but a probability distribution, with multiple network instances forming an ensemble (Lakshminarayanan et al., 2017). In related work, a similar approach was recently applied for age estimation from fetal brain MRI, reporting high accuracy and promising indications for abnormality detection (Shi et al., 2020).

The aim of this work is to develop an automated strategy for body composition analysis on UK Biobank neck-to-knee body MRI which provides not only measurements (Langner et al., 2020b) but also introduces individual uncertainty estimates that can represent confidence intervals. As a key advantage, the deep regression approach can be trained without access to reference segmentation masks and instead learns to emulate the existing, numerical metadata. Six body composition measurements relating to adipose tissues with high relevance for cardiometabolic disease were predicted from two-dimensional representations of the MRI data. ResNet50 neural network instances (He et al., 2016) for image-based regression were trained to each predict the mean and variance of a Gaussian probability distribution over a given measurement value. Combined into ensembles they provided estimates of predictive uncertainty (Lakshminarayanan et al., 2017). The main contribution consists in extensive analysis of the independent effects of *mean-variance* regression and ensembling on overall accuracy and speed, but also on the calibration (Guo et al., 2017) of uncertainties and their ability to identify the worst predictions in sparsification (Ilg et al., 2018), both in cross-validation on about 8500 subjects and testing on another 1000 subjects. The proposed method was deployed for inference to obtain previously unavailable measurements from more than 30,000 images, including 1000 repeat scans.

2. Materials and methods

The neck-to-knee body MRI of each subject was formatted into a two-dimensional image from which the proposed method estimates a numerical measurement value in image-based regression. This work examines *least squares* regression, which produces only the measurement value itself, (Langner et al., 2020b,c), but also *mean-variance* regression (Nix and Weigend, 1994), in which both the mean value and the variance of a Gaussian probability distribution over one measurement of one subject is modeled. In *ensembling*, the predictions of several networks are furthermore aggregated (Lakshminarayanan et al., 2017). The thus obtained uncertainty estimates can help to identify outliers and potential failure cases automatically (Gustafsson et al., 2020).

2.1. UK Biobank image data

UK Biobank has recruited more than half a million men and women by letter from the National Health Service in the United Kingdom, starting in 2006 (Sudlow et al., 2015). Examinations involve several visits to UK Biobank assessment centers, with imaging procedures launching in 2014 for a subgroup of 100,000 participants (Littlejohns et al., 2020). At the time of writing, medical imaging data from three different centers has been released for 40,264 men and women (52% female) aged 44–82 (mean 64) years with BMI 14–62 (mean 27) kg/m^2 and a majority of 94% with self-reported White-British ethnicity.

For 1209 of these, data from a repeat imaging visit with an offset of about two years has been released. All participants provided informed consent and both the UK Biobank examinations and the experiments in this work were approved by the responsible British and Swedish ethics committees.

2.1.1. MRI data

The MRI protocol examined in this work is listed as UK Biobank field 20201 and covers the body from neck to knee in six separate imaging stations acquired in a scan time below ten minutes (West et al., 2016; Littlejohns et al., 2020). Volumetric, co-aligned images of water and fat signal were acquired with a two-point Dixon technique with TR = 6.69, TE = 2.39/4.77 ms and flip angle 10 deg on a Siemens Aera Magnetom 1.5 device. The image resolution varies between stations, with a typical grid of (224 × 174 × 44) voxels of (2.232 × 2.232 × 4.5) mm (for more detail, see "Body MRI protocol parameters" in Littlejohns et al. (2020)).

2.1.2. Image formatting

For this work, the six MRI stations of each subject were first fused into a common voxel grid by trilinear interpolation to form a single volume of (224 × 174 × 370) voxels for each signal type. These volumes were then converted to two-dimensional representations by summing all values along two axes of view, yielding a coronal and sagittal mean intensity projection, which were concatenated side by side. This was done separately for both the water and fat signal, with the resulting images individually normalized and downsampled to form two color channels of a single image of (256 × 256 × 2) pixels (Langner et al., 2020b). As a third image channel, both a single coronal and sagittal fat fraction slice were extracted based on a body mask (Langner et al., 2020c). These fractions resulted from voxel-wise division of the fat signal by the sum of water and fat signal. Fig. 1 shows the result, a dual mean intensity projection with fat fraction slices, encoded in 8 bit for faster processing.

2.2. Ground truth

UK Biobank provides several body composition measurements from the same neck-to-knee body MRI data as used in this work, based on volumetric multi-atlas segmentations (West et al., 2016; Borga et al., 2015): Visceral Adipose Tissue (VAT), abdominal Subcutaneous Adipose Tissue (SAT), Total Adipose Tissue (TAT), Total Lean Tissue (TLT), and

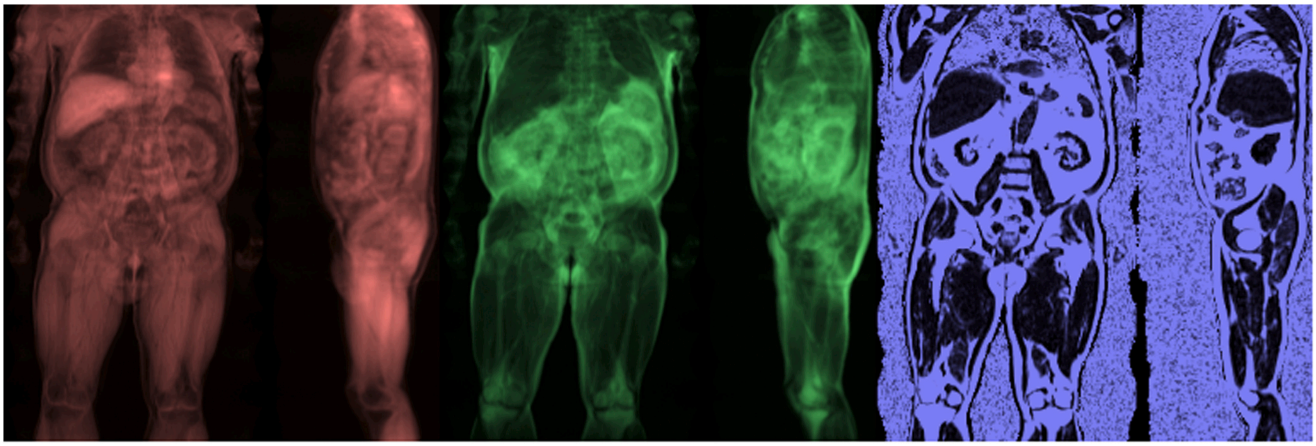


Fig. 1. As input to the neural network, each MRI volume was represented as color image of (256 × 256 × 3) pixels by forming channels from the projected water (red) and fat (green) signal and fat fraction slices (blue) from two axes each.

Total Thigh Muscle (TTM). Together with Liver Fat Fraction (LFF) values based on dedicated multi-echo liver MRI (Linge et al., 2018), these reference measurements form the ground truth data, or regression targets, for this work.

2.3. Data partitions

Among the 40,264 released images of the initial imaging visit, visual inspection identified 1376 subjects with artifacts such as water-fat signal swaps, non-standard positioning and metal objects (Langner et al., 2020b). Three datasets were formed from the initial imaging visit from those subjects for whom any of the six reference measurements were available.

Dataset D_{cv} consists of 8539 subjects without artifacts and was subdivided into a 10-fold cross-validation split which was retained for all experiments.

Dataset D_{test} contains another 1107 subjects without artifacts and served as a test set, but notably lacks any values for two of the six regression targets for which no reference values have been released yet.

Dataset D_{art} was formed from those subjects with identified artifacts, yielding 330 subjects, to examine behavior on abnormal data.

Two additional datasets were formed from those subjects with no available reference measurements. Dataset D_{infer} comprises all remaining 29,234 subjects without artifacts from the initial imaging visit, for whom the prediction model was applied to for inference. Finally, dataset $D_{revisit}$ was formed for inference on the repeat imaging visit from 1179 subjects with no image artifacts.

2.4. Model

A ResNet50 architecture (He et al., 2016) was configured to receive the two-dimensional image format as seen in Fig. 1 as input for a given subject and predict all six regression targets at once. No explicit segmentation was performed at any stage of this work. Each network was pre-trained on ImageNet and optimized with Adam (Kingma and Ba, 2014) at batch size 32 with online augmentation by random translations. After 5000 iterations, the base learning rate of 0.0001 was reduced by factor 10 and training continued for another 1000 iterations (Langner et al., 2020b). All experiments were conducted in PyTorch, using an Nvidia RTX 2080 Ti graphics card with 11 GB RAM.

Four distinct configurations were compared. As the first one, a *least squares* regression network predicted only these six output values, each corresponding to one measurement for a given subject, trained by optimizing the mean squared error criterion of equation 1. In this formula, $\mu_\theta(\mathbf{x}_n)$ represents the network prediction for the n-th input sample $\mathbf{x}_n$, with y_n as the corresponding ground truth value.

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \mu_\theta(\mathbf{x}_n))^2 \tag{1}$$

As a second configuration, *least squares ensembles* were formed by combining ten such networks. Their predictions were averaged and the spread, or empirical variance, of their predictions used as uncertainty estimate (Ilg et al., 2018).

As the third configuration, *mean-variance* regression was performed by predicting two values, corresponding to the mean and variance of a Gaussian probability distribution over one measurement value for a given subject, optimized with a negative log-likelihood criterion (Nix and Weigend, 1994) as shown in equation 2. Here, $p_\theta(y_n|\mathbf{x}_n)$ is the probabilistic predictive distribution over one measurement value, modeled by the network outputs $\mu_\theta(\mathbf{x}_n)$ and $\sigma^2_\theta(\mathbf{x}_n)$, which represent the predicted mean and corresponding predicted variance for input sample $\mathbf{x}_n$, respectively. The last term, c, is a constant that does not depend on θ. This criterion expands the mean squared error of eq. 1 by a sample-specific, heteroscedastic variance and can likewise be averaged across multiple samples. This predicted variance directly serves as an estimate of uncertainty, with high values describing a wide normal distribution within which plausible values for the estimated measurement are assumed.

$$-\log p_\theta(y_n|\mathbf{x}_n) = \frac{\log\sigma^2_\theta(\mathbf{x}_n)}{2} + \frac{(y_n - \mu_\theta(\mathbf{x}_n))^2}{2\sigma^2_\theta(\mathbf{x}_n)} + c \tag{2}$$

As the fourth and final configuration, *mean-variance ensembles* employ ten such network instances. Their predictions can likewise be aggregated to obtain estimates of predictive uncertainty (Lakshminarayanan et al., 2017).

In all ensembles, model diversity was increased by withholding one of ten evenly sized subsets of the training data from each instance, as if they had been obtained from a preceding cross-validation experiment. The target values were standardized (Langner et al., 2020b). When one or more of the six ground truth values for a given training sample were missing, their contribution to the loss term was dynamically set to zero, so that they would not affect the training process. In this way, it was possible to utilize samples with missing values and provide as much training data as possible. A PyTorch implementation for training and inference will be made publicly available.[1]

[1] github.com/tarolangner/ukb_mimir

2.5. Evaluation

All configurations were evaluated in 10-fold cross-validation on dataset D_{cv} and also validated against artifact dataset D_{art}. The best configuration was eventually applied to test dataset D_{test} and deployed for inference on datasets D_{infer} and $D_{revisit}$.

The predicted measurements were compared to the reference values with the intraclass correlation coefficient (ICC) with a two-way random, single measures, absolute agreement definition (Koo and Li, 2016) and the coefficient of determination R^2. The mean absolute error (MAE) is also reported, together with the mean absolute percentage error (MAPE) as a relative error measurement. The latter is the absolute difference between prediction and reference divided by the reference. Additionally, aggregated saliency maps were generated to highlight relevant image areas (Selvaraju et al., 2017).

The estimated uncertainties were evaluated regarding sparsification (Ilg et al., 2018) and calibration (Guo et al., 2017). Sparsification examines whether the highest uncertainties coincide with the highest prediction errors. Ranking all measurements by their uncertainty and excluding one after another should accordingly yield consistent improvements in performance metrics such as the MAE. Calibration examines the magnitude of uncertainties and resulting under- or overconfidence of predictions. The uncertainty obtained for any given sample corresponds to the variance of a Gaussian probability distribution, modeling characteristic confidence intervals around the predicted mean. Higher uncertainty scales these intervals to be wider, enabling them to cover larger errors. Ideally calibrated uncertainties define confidence intervals that cover, on a set of samples, a percentage of errors that corresponds exactly to their specific confidence level.

3. Results

Both *mean-variance* regression and ensembling provided complementary benefits. Combining both yielded the best predictive performance, shown in Table 1 and Fig. 2, with additional detail provided in the supplementary material. On average, the predictions can account for 98% (R^2) of the variability in reference values, with absolute agreement (ICC) above 0.97 on all targets. The metrics carry over to the test data largely unchanged. All targets are predicted with a relative error below 5%, except the liver fat fraction. This target also incurred the highest relative uncertainties and is examined further in the supplementary material, together with additional evaluation metrics, and a comparison to alternative reference methods. It also provides additional detail on the saliency analysis, which is compiled into Fig. 3.

Fig. 4 shows that even without utilizing the uncertainties, the *mean-variance* regression ensemble reduces the MAE by 12% when compared to the *least-squares* regression baseline. The uncertainties enable sparsification, identifying some of the worst predictions which can be excluded to reduce the prediction error even further. The scatter plots of Fig. 2 show predictions for one target in detail, together with color-coded uncertainty. Despite containing image artifacts, not all subjects of dataset D_{art} yield higher uncertainties than the normal material. Indeed, many of these subjects result in highly accurate predictions despite the artifacts, and high uncertainties tend to occur only in those cases with high prediction errors. On test dataset D_{test}, the uncertainty highlights an outlier case for VAT (see Fig. 2), SAT, and TTM. This one subject causes consistently flawed predictions and was found to suffer from an abnormal, atrophied right leg.

Table 1
Evaluation results.

Target name		Cross-Validation		Testing	
		ICC	% error	ICC	% error
Visceral Adipose Tissue	(VAT)	0.997	4.2	0.997	3.6
Abdominal Subcutaneous Adipose Tissue	(SAT)	0.996	2.8	0.996	2.7
Total Adipose Tissue	(TAT)	0.997	1.8	/	/
Total Lean Tissue	(TLT)	0.983	2.5	/	/
Total Thigh Muscle	(TTM)	0.996	1.6	0.995	1.6
Liver Fat Fraction	(LFF)	0.979	25.7	0.982	21.6

*Results for the *mean-variance* ensemble on cross-validation dataset D_{cv} and testing on dataset D_{test}, with intraclass correlation coefficient (ICC) and MAPE (% error).

On datasets D_{cv} and D_{test} the predicted means exhibit a consistent, linear correlation with the predicted log uncertainties. Accordingly, large subjects with high volumes induce systematically higher uncertainty. Although these cases also generally incur higher prediction errors, this bias can be shown to not achieve optimal sparsification. On the normal material with hardly any outliers, this tendency is so strong that sparsifying simply by predicted mean is almost as effective as using the uncertainties. On dataset D_{art}, this bias is less pronounced, as those cases with artifacts that cause genuine prediction failures are correctly assigned much higher uncertainty.

The best calibration was also achieved by the *mean-variance* ensemble, which nonetheless often produced overconfident uncertainties. Post-processing with target-wise scaling factors can achieve a near perfect fit to the validation data, however, and also improves the overall calibration on the test set. The supplementary material explores both sparsification and calibration in more detail and also lists results for datasets D_{infer} and $D_{revisit}$, on which the proposed method inferred new measurements for over 30,000 images.

No difference in processing speed was observed between *least squares* and *mean-variance* regression. Image formatting required the bulk of processing time, but once cached, training one network only requires about 15 min, or 2.5 h for an ensemble of ten instances. Ensemble predictions for about 60 subjects can be generated within one second, so that inference for all 30,000 required less than ten minutes.

4. Discussion

With relative measurement errors below 5%, all targets except the liver fat fraction can be predicted with higher accuracy than observed for the mutual agreement between the reference and alternative established methods, both in cross-validation and on the test data. For liver fat itself, the relative error of 22–26% is worse than the 15% seen between the reference used here and an alternative set of UK Biobank liver fat measurements. The two-point Dixon images inherently limit the prediction accuracy for this target, as the reference values were obtained from another imaging protocol that reconstructs fat fractions more faithfully (Wilman et al., 2017; Linge et al., 2018). The saliency analysis of Fig. 3 indicates that the networks nonetheless learned to correctly identify liver tissue and other target-specific regions. The inference on 30,000 subjects provides material for further medical study which is, however, beyond the scope of this work.

The estimated uncertainties identified many of the worst prediction errors. They correctly highlighted an outlier with abnormal physiology on the test data and enabled consistent reductions in the mean prediction error by excluding the least certain measurements. On the inference datasets, the highest uncertainties were furthermore found in several cases to coincide with previously undetected anomalies in positioning, but also with minor artifacts and pathologies that may have negatively affected prediction accuracy and should arguably have been excluded during the original quality controls. In practice, the acquired measurements can accordingly be supplied together with their uncertainty, which could serve both as an error estimate and as a means to identify potential anomalies and failure cases. The affected cases could then be manually examined and, if necessary, excluded from further analyses.

However, the results also show two noteworthy limitations of the proposed approach which arise from imperfect calibration and the observed bias for high measurement values to incur high uncertainties. The imperfect calibration is linked to uncertainties that often

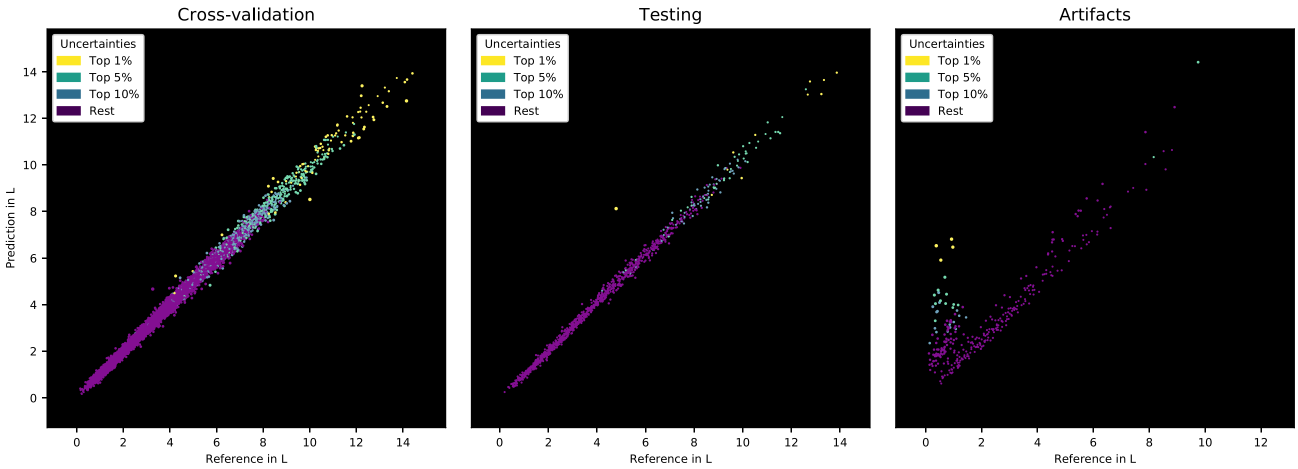


Fig. 2. Mean-variance ensemble predictions and reference values for Visceral Adipose Tissue (VAT) in cross-validation on D_{cv}, testing on D_{test}, and on subjects with artifacts of D_{art}, depicted with color-coded uncertainty. The listed percentiles refer to those samples with the highest uncertainty.

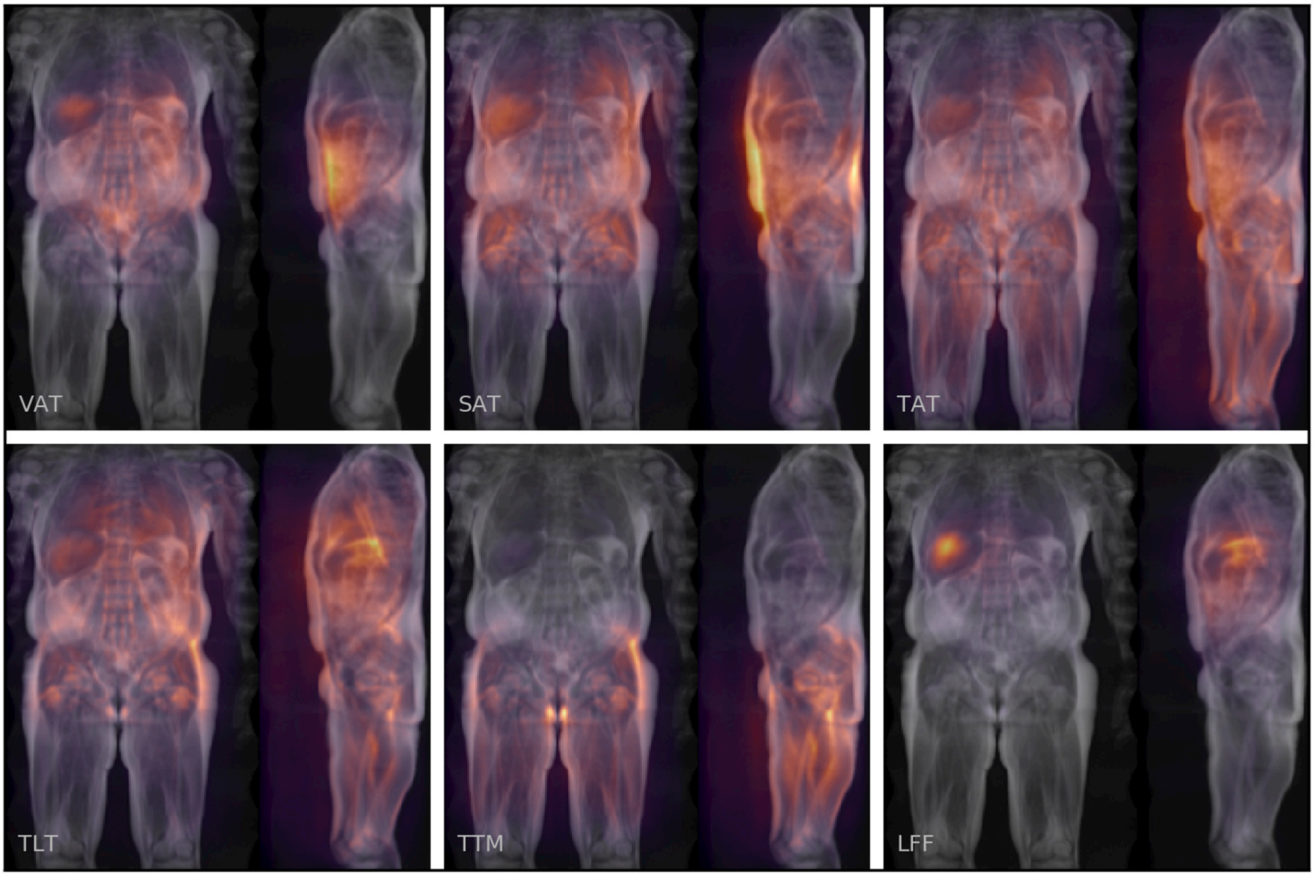


Fig. 3. Co-registered, aggregated saliency information for about 3000 subjects, showing the fat signal channel only (see supplementary material for more).

underestimate the true measurement error. This is a known effect related to overfitting on the training data (Guo et al., 2017; Laves et al., 2020). As shown in the supplementary material, it is possible to correct the calibration by calculating target-wise scaling factors on the validation results. Once obtained, these simple scaling factors also yield improved overall calibration on the test data.

The bias towards systematically higher uncertainty in higher measurement values is a more concerning pattern. This effect can make it hard to distinguish whether a measurement with high uncertainty should be excluded due to being flawed or whether it merely resulted from a large subject, many of whom may provide valuable insight in correlation studies. It is most pronounced in the normal material where no genuine failure cases are encountered. In contrast, the uncertainty for one abnormal subject in the test set or the flawed predictions on images with artifacts of dataset D_{art} are typically higher.

Conceptually, body weights above 150 kg and BMIs of up to 53 kg/m^2 as present in the training data represent physiological extremes that could be considered outliers in their own right. Arguably, the two-dimensional projections are also inherently less suitable to represent more voluminous bodies and many of the largest subjects furthermore

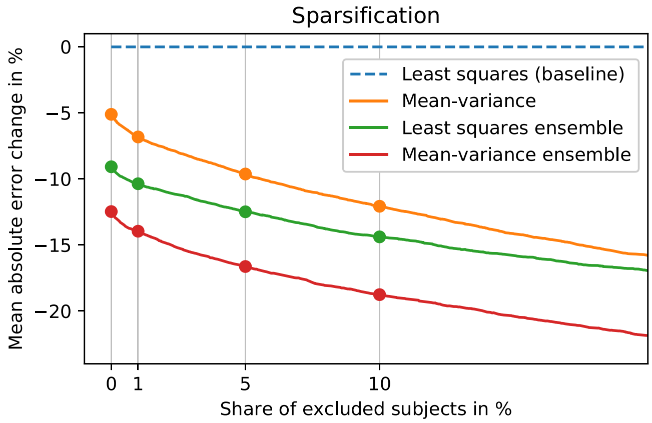


Fig. 4. This sparsification plot (Ilg et al., 2018) shows how the overall performance can be improved by gradually excluding those subjects with the highest predicted measurement uncertainty. Each position along the x-axis represents a certain share of excluded, most uncertain measurements, whereas the y-axis shows the change in mean absolute error relative to baseline, averaged across all targets on dataset D_{cv}. Even without utilizing the uncertainty to exclude any subjects, the *mean-variance* ensemble achieves a reduction of the MAE by 12%. Further improvements in the MAE can be achieved excluding increasingly large shares of those measurements with highest uncertainty.

show considerable variability in shape and extend beyond the field of view. Even then, the effect is gradual and large subjects incur higher uncertainty than warranted in terms of the prediction errors alone. Previous work on age estimation from fetal brain MRI reported similar effects (Shi et al., 2020), noting specifically that higher aleatoric uncertainty, corresponding to the variances returned by the network instances, correlated with higher gestational age of the fetal brain. In this work, the effect is present in both the aleatoric and epistemic uncertainty component as modeled by the empirical variance, even in *least-squares* regression ensembles. On a technical level, the *mean--variance* configuration provided immediate benefits over *least squares* regression despite merely changing the loss function and requiring that both a mean and a variance be predicted. This could be explained by *loss attenuation* (Kendall and Gal, 2017; Ilg et al., 2018) weakening the impact of outliers among the ground truth values. Several mismatches between the image data and reference were identified where the predictions also incur high errors in spite of low uncertainty. Images with artifacts, in contrast, did not necessarily yield high uncertainties, as the method was in fact able to provide accurate predictions for many of them. In turn, this also means that subjects with artifacts will not generally be identified as out-of-distribution samples. Ensembling yielded an inherent benefit in prediction accuracy and also improved the calibration. The ten network instances were conveniently obtained from a cross-validation split, but sufficient ensemble diversity could potentially be induced by random weight initialization alone and similar benefits can be achieved with fewer instances as seen in ablation experiments of the supplementary material and related literature (Fort et al., 2019; Ovadia et al., 2019). Based on the results, even a single *mean-variance* instance would be viable in practical settings if model size and runtime are of chief concern. The calibration could be adjusted with scaling factors, although it would not benefit from the 12% reduction in MAE achieved by ensembling.

Several additional limitations apply on a methodological level. No independent, external test set was examined, so that no claim can be made about generalization of the trained networks to other studies. The validation and test cases used in this work are furthermore preselected for the intended measurements by virtue of having passed the quality controls of the reference methods. Similarly, certain phenotypes were systematically excluded from the experiments in this paper, such as subjects with knee implants or other severe pathologies. When applied to different imaging devices, protocols, or subject demographics, new training data in the range of several hundred samples would likely be required. In contrast, multi-atlas segmentations with manual corrections have been based on just above 30 annotated subjects (West et al., 2016), whereas neural networks for semantic segmentation typically report training data ranging from 90 to 220 subjects (Fitzpatrick et al., 2020; Bagur et al., 2020) on UK Biobank MRI.

When compared to neural networks for segmentation, the proposed approach accordingly requires more training samples and produces no output segmentation masks. In turn, it can be trained without access to reference segmentations in an end-to-end fashion that does not require for the property of interest to be manually encoded in the input data during training. Previous work showed that it outperformed segmentation in estimating liver fat from the two-point Dixon images, possibly by using additional image information that is not easily accessible to human intuition Langner et al. (2020c), and also accurately estimated other, more abstract properties Langner et al. (2020b). Likewise, the uncertainty quantification as proposed here can provide error bounds for the measurement that is ultimately of interest for medical research, although approaches for voxel-wise uncertainty from segmentation networks have also been proposed in the literature Roy et al. (2019).

The concept of designing two-dimensional input formats resembles hand-crafted feature selection and it would be preferable to apply a regression technique directly to the volumetric MRI data. No claim is intended for the chosen representation to be optimal as input to the neural network. The MRI volumes could be sliced, projected, or aggregated in various ways and in any signal or phase component may contain valuable information. Despite the empirical success of the presented approach, further improvements may be possible, as the chosen format compresses the MRI data to just 0.5% of its original size and almost certainly results in a loss of information. However, a fully volumetric approach would likely require substantially increased processing time and GPU memory. The proposed approach, in contrast, can run on consumer-grade hardware and achieves relative errors as low as 1.6%, which may be hard to improve much further. Future work may adapt the presented approach to the dedicated liver MRI of UK Biobank, with potential for far more accurate liver fat predictions.

Future work may also explore how the bias between high measurements and high uncertainty can be corrected for and could explore alternative strategies which are known to produce substantially distinct estimates of uncertainty (Stahl et al., 2020). However, it is unclear whether Monte-Carlo techniques that employ dropout at test time (Gal and Ghahramani, 2016) could reach sufficient predictive performance, whereas more faithful approximations of Bayesian inference with Markov chain Monte-Carlo (Neal, 2012) may not be computationally viable. Deep ensembles are often reported as one of the most successful strategies (Gustafsson et al., 2020; Ovadia et al., 2019; Ashukha et al., 2020) and a suitable alternative will have to achieve better calibration and sparsification without sacrificing predictive accuracy or exceeding the computational limitations in order to be competitive.

In a large-scale study such as the UK Biobank the main strengths of the proposed approach can be exploited. Without any need for further guidance, corrections, or intervention, these values can be inferred for the entire imaged study population, both for existing and future imaging data. The resulting measurements can be obtained for further study and quality control months or years before full coverage has been achieved with the reference techniques. In practice, researchers may apply this system to obtain automated measurements for all upcoming 120,000 UK Biobank neck-to-knee body MRI scans yet to be released, and will be alerted to potential prediction failures by the predictive uncertainty. Future developments may also yield comparable systems that could ultimately be integrated into scanner software to provide fully automated analyses for specific imaging protocols.

5. Conclusion

In conclusion, both *mean-variance* regression and ensembling provided complementary benefits for the presented task. Without extensive architectural changes or prohibitive increases in computational cost they enabled fast and accurate measurements of body composition for the entire imaged UK Biobank cohort. The predicted uncertainty can, despite the specified limitations, give valuable insight into potential failure cases and will be made available together with the inferred measurements for further medical studies.

CRediT authorship contribution statement

T.L. wrote the manuscript, helped devise and conducted the experiments. F.K.G. and B.A. helped devise the experiments. R.S. and H.A. helped revise the manuscript. J.K. obtained the data and helped revise the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by a research grant from the Swedish Heart- Lung Foundation and the Swedish Research Council (2016-01040, 2019-04756, 2020-0500, 2021-70492) and used the UK Biobank Resource under application no. 14237.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compmedimag.2021.101994.

References

Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D., 2020.Pitfalls of in-domain uncertainty estimation and ensembling in deep learning.In: International Conference on Learning Representations.⟨https://openreview.net/forum?id=BJxI5gHKDr⟩.

Bagur, A.T., Ridgway, G., McGonigle, J., Brady, M., Bulte, D., 2020.Pancreas segmentation-derived biomarkers: Volume and shape metrics in the uk biobank imaging study.In: Annual Conference on Medical Image Understanding and Analysis. Springer, 131–142.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J. Cardiovasc. Magn. Reson. 20 (1), 65.

Basty, N., Liu, Y., Cule, M., Thomas, E.L., Bell, J.D., Whitcher, B., 2020. Automated measurement of pancreatic fat and iron concentration using multi-echo and t1-weighted mri data. 2020 IEEE 17th International Symposium on Biomedical Imaging ((ISBI)), pp. 345–348.

Borga, M., 2018. Mri adipose tissue and muscle composition analysis-a review of automation techniques. Br. J. Radiol. 91 (1089), 20180252.

Borga, M., Thomas, E.L., Romu, T., Rosander, J., Fitzpatrick, J., DahlqvistLeinhard, O., Bell, J.D., 2015. Validation of a fast method for quantification of intra-abdominal and subcutaneous adipose tissue for large-scale human studies. NMR Biomed. 28 (12), 1747–1753.

Cole, J.H., Ritchie, S.J., Bastin, M.E., Hernández, M.V., Maniega, S.M., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., et al., 2018. Brain age predicts mortality. Mol. Psychiatry 23 (5), 1385–1392.

Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M.M., Reuter, M., 2020. Fatsegnet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon mri. Magn. Reson. Med. 83 (4), 1471–1483.

Fitzpatrick, J., Basty, N., Cule, M., Liu, Y., Bell, J.D., Thomas, E.L., Whitcher, B., 2020. Large-scale analysis of iliopsoas muscle volumes in the uk biobank. arXiv: ⟨http://arXiv.org/abs/arXiv:2008.05217⟩.

Fort, S., Hu, H., Lakshminarayanan, B., 2019.Deep ensembles: A loss landscape perspective. arXiv: ⟨http://arXiv.org/abs/arXiv:1912.02757⟩.

Gal, Y., Ghahramani, Z., 2016.Dropout as a bayesian approximation: Representing model uncertainty in deep learning.In: international conference on machine learning.1050–1059.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521 (7553), 452–459.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017.On calibration of modern neural networks. arXiv: ⟨http://arXiv.org/abs/arXiv:1706.04599⟩.

Gustafsson, F.K., Danelljan, M., Bhat, G., Schön, T.B., 2020a.Energy-based models for deep probabilistic regression.In: European Conference on Computer Vision.Springer, 325–343.

Gustafsson, F.K., Danelljan, M., Schon, T.B., 2020b.Evaluating scalable bayesian deep learning methods for robust computer vision.In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.318–319.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition ((CVPR)), pp. 770–778.

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T., 2018. Uncertainty estimates and multi-hypotheses networks for optical flow.In: Proceedings of the European Conference on Computer Vision (ECCV).pp.652–667.

Irving, B., Hutton, C., Dennis, A., Vikal, S., Mavar, M., Kelly, M., Brady, J.M., 2017.Deep quantitative liver segmentation and vessel exclusion to assist in liver assessment.In: Annual Conference on Medical Image Understanding and Analysis.Springer, 663–673.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 5574–5584.

Kingma, D.P., Ba, J., 2014.Adam: A method for stochastic optimization. arXiv: ⟨http://arXiv.org/abs/arXiv:1412.6980⟩.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropractic Med. 15 (2), 155–163.

Küstner, T., Hepp, T., Fischer, M., Schwartz, M., Fritsche, A., Häring, H.-U., Nikolaou, K., Bamberg, F., Yang, B., Schick, F., et al., 2020. Fully automated and standardized segmentation of adipose tissue compartments via deep learning in 3d whole-body mri of epidemiologic cohort studies. Radiology 2 (5), e200010.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 6402–6413.

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

Langner, T., Östling, A., Maldonis, L., Karlsson, A., Olmo, D., Lindgren, D., Wallin, A., Lundin, L., Strand, R., Ahlström, H., et al., 2020a. Kidney segmentation in neck-to-knee body mri of 40,000 uk biobank participants. Sci. Rep. 10 (1), 1–10.

Langner, T., Strand, R., Ahlström, H., Kullberg, J., 2020b. Large-scale biometry with interpretable neural network regression on uk biobank body mri. Sci. Rep. 10 (1), 1–9.

Langner, T., Strand, R., Ahlström, H., Kullberg, J., 2020c.Large-scale inference of liver fat with neural networks on uk biobank body mri.In: International Conference on Medical Image Computing and Computer-Assisted Intervention.Springer, 602–611.

Laves, M.-H., Ihler, S., Fast, J.F., Kahrs, L.A., Ortmaier, T., 2020. Well-calibrated regression uncertainty in medical imaging with deep learning. Med. Imaging Deep Learn.

Linge, J., Borga, M., West, J., Tuthill, T., Miller, M.R., Dumitriu, A., Thomas, E.L., Romu, T., Tunón, P., Bell, J.D., et al., 2018. Body composition profiling in the uk biobank imaging study. Obesity 26 (11), 1785–1795.

Littlejohns, T.J., Holliday, J., Gibson, L.M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J.D., Boultwood, C., Collins, R., Conroy, M.C., et al., 2020. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. 11 (1), 1–12.

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

Neal, R.M., 2012. Bayesian Learning for Neural Networks, 118. Springer Science & Business Media,.

Nix, D.A., Weigend, A.S., 1994.Estimating the mean and variance of the target probability distribution.In: Proceedings of 1994 ieee international conference on neural networks (ICNN'94), 1.IEEE, 55–60.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your modelas uncertainty? evaluating predictive uncertainty under dataset shift. Adv. Neural Inf. Process. Syst. 13991–14002.

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2 (3), 158.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195, 11–22.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., Oct. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, 618–626. ‘ ⟨http://ieeexplore.ieee.org/document/8237336/⟩.

Shi, W., Yan, G., Li, Y., Li, H., Liu, T., Sun, C., Wang, G., Zhang, Y., Zou, Y., Wu, D., 2020. Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. NeuroImage 223, 117316.

Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., 2020.Evaluation of uncertainty quantification in deep learning.In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems.Springer, 556–568.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., . 2015. 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. ⟨https://dx.plos.org/10.1371/journal.pmed.1001779⟩.

Wang, Y., Qiu, Y., Thai, T., Moore, K., Liu, H., Zheng, B., 2017. 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. 144, 97–104.

West, J., Dahlqvist Leinhard, O., Romu, T., Collins, R., Garratt, S., Bell, J.D., Borga, M., Thomas, L., . 2016. Feasibility of MR-based body composition analysis in large scale population studies. PLoS One 11, 9. https://doi.org/10.1371/journal.pone.0163332. ⟨https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5035023/⟩.

Weston, A.D., Korfiatis, P., Kline, T.L., Philbrick, K.A., Kostandy, P., Sakinis, T., Sugimoto, M., Takahashi, N., Erickson, B.J., 2019. Automated abdominal segmentation of ct scans for body composition analysis using deep learning. Radiology 290 (3), 669–679.

Wilman, H.R., Kelly, M., Garratt, S., Matthews, P.M., Milanesi, M., Herlihy, A., Gyngell, M., Neubauer, S., Bell, J.D., Banerjee, R., et al., 2017. Characterisation of liver fat in the uk biobank cohort. PLoS One 12 (2), e0172921.

Xue, W., Islam, A., Bhaduri, M., Li, S., 2017. Direct multitype cardiac indices estimation via joint representation and regression learning. IEEE Trans. Med. Imaging 36 (10), 2057–2067.