Large-scale inference of liver fat with neural networks on UK Biobank body MRI[*]

Taro Langner[1], Robin Strand[1], Håkan Ahlström[1,2], and Joel Kullberg[1,2]

[1] Uppsala University, 751 85 Uppsala, Sweden
`taro.langner@surgsci.uu.se`
[2] Antaros Medical, BioVenture Hub, 431 53 Mölndal, Sweden

Abstract. The UK Biobank Imaging Study has acquired medical scans of more than 40,000 volunteer participants. The resulting wealth of anatomical information has been made available for research, together with extensive metadata including measurements of liver fat. These values play an important role in metabolic disease, but are only available for a minority of imaged subjects as their collection requires the careful work of image analysts on dedicated liver MRI. Another UK Biobank protocol is neck-to-knee body MRI for analysis of body composition. The resulting volumes can also quantify fat fractions, even though they were reconstructed with a two- instead of a three-point Dixon technique. In this work, a novel framework for automated inference of liver fat from UK Biobank neck-to-knee body MRI is proposed. A ResNet50 was trained for regression on two-dimensional slices from these scans and the reference values as target, without any need for ground truth segmentations. Once trained, it performs fast, objective, and fully automated predictions that require no manual intervention. On the given data, it closely emulates the reference method, reaching a level of agreement comparable to different gold standard techniques. The network learned to rectify non-linearities in the fat fraction values and identified several outliers in the reference. It outperformed a multi-atlas segmentation baseline and inferred new estimates for all imaged subjects lacking reference values, expanding the total number of liver fat measurements by factor six.

Keywords: Magnetic resonance imaging (MRI) · Liver Fat · Neural network.

---

1   Introduction

The UK Biobank has recruited more than 500,000 volunteer participants for medical examination, 100,000 of whom are planned to undergo extensive imaging procedures [14]. Several modalities are involved, including dedicated MRI of the brain, heart, pancreas, and liver, the latter of which enables measurements of accumulated liver fat [16], closely linked to type 2 diabetes and other metabolic disorders. A fat content of 5.5% has been defined as threshold for non-alcoholic fatty liver disease (NAFLD) [1], which can progress to non-alcoholic steatohepatitis (NASH) [11] and eventually to liver cirrhosis with potentially fatal outcome [19]. Both MRI and magnetic resonance spectroscopy (MRS) are non-invasive alternatives to liver biopsy for reliable quantification [13].

Due to its scale, the UK Biobank has the potential to relate liver fat as a biomarker to the wide range of metadata, such as disease outcomes, life-style factors and genetic profiles. At the time of writing, almost 40,000 participants have completed the UK Biobank imaging procedures, but only about 5,000 reference liver fat measurements are available, based on transverse slices acquired from the dedicated liver MRI [16]. Like these images, the volumes acquired by neck-to-knee body MRI for body composition analysis [15] also include the liver and encode voxel-wise proton density fat fractions (PDFF). Due to using a two- instead of a three-point Dixon technique for reconstruction, these images may encode systematically different PDFF values, and similar protocols have previously shown low agreement with other established methods for quantification of liver fat [8]. However, the UK Biobank has released more than 30,000 neck-to-knee body MRI scans, which can be evaluated with machine learning techniques.

Various biological metrics, including liver fat, can be automatically inferred on image data from these scans by convolutional neural networks for regression [9]. Similar strategies have been previously applied to a range of medical imaging modalities including MRI, with the goal of quantifying properties such as age [2][4][10] and structures of the heart [17], but also blood pressure and smoking status [12]. This technique is distinct from neural networks trained for segmentation, which can also be applied for liver fat measurements [7], but typically require ground truth segmentations for training.

In this work, liver fat measurements are inferred by automated analysis of the more readily available neck-to-knee body MRI with neural networks trained for regression on data from these images [15] and the UK Biobank reference values [16] as ground truth. The following contributions are made. (1) A specialized framework for liver fat inference from UK Biobank neck-to-knee body MRI, which adapts a generalized system [9] for superior performance. (2) A three-way comparison between this method, the reference, and a simple multi-atlas segmentation baseline. (3) Inferred liver fat values for more than 30,000 UK Biobank subjects, which could be shared for medical research and as a baseline for quality control. Code samples and documentation for these implementations are publicly available.[3]

---

[3] https://github.com/tarolangner/mri-biometry

2 Methodology

2.1 UK Biobank Data

UK Biobank participants were recruited by letter from the National Health Service and scanned at three different imaging centers in the United Kingdom. The majority of subjects reported white British ethnicity ($\sim$ 94%) with a mean age of 64 years (range 44-82, standard deviation 7.5), mean BMI of 26.6 kg/m^2 (standard deviation 4.3) and a share of 52% males.

Reference Liver Fat Measurements Reference liver fat measurements were available for 4,613 subjects as field 22402-2.0 of the UK Biobank [16]. These values are based on the PDFF map of a single transverse slice of the liver, acquired with a Siemens 1.5T MAGNETOM Aera and three-point Dixon technique. The reference method returns the mean PDFF of three manually placed ROIs with liver tissue, avoiding vessels and other confounding structures. To avoid confusion with relative values, liver fat percentage points are referred to as fat fractions (FF). The available values range from 0-46 FF, with a mean of 3.9 $\pm$ 4.6 FF (median 2.1 FF). Of the total 4,613 subjects, 920 (20%) have recorded liver fat values above 5.5 FF.

Neck-to-knee Body MRI The UK Biobank neck-to-knee body MRI scans released as field 20201-2.0 were also acquired with a Siemens 1.5T MAGNETOM Aera device, using a dual-echo Dixon technique with TR=6.69 and TE=2.39/4.77 ms, and flip angle 10deg[15]. The resulting water-fat volumes cover most of the body with six overlapping stations, excluding the head and lower legs, whereas the arms and other tissues at the edges of the magnetic field are typically subject to strong image artifacts. Likewise, the borders between stations often contain motion artifacts that can affect the shape of the liver.

2.2 Experimental Setup

A neural network was trained for regression of liver fat values on the neck-to-knee body MRI of those subjects with available reference values. It was first evaluated in 10-fold cross-validation and then, after training on the full dataset, applied to the remaining subjects for inference. In both phases, independent measurements from a simple multi-atlas segmentation strategy described in one of the following sections served as a baseline.

Datasets All 32,323 neck-to-knee body MRI scans that have been released at the time of writing were quality controlled by an operator who visually inspected two-dimensional mean intensity projections of the water and fat signals [9]. Due to water-fat swaps, noise, metal objects, unusual positioning and other artifacts, about 3,6% of the subjects were excluded, leaving 31,171 images for the experiments.

Three datasets were formed from those subjects that passed the quality control. Validation dataset A consists of those 4,418 subjects with existing reference values as ground truth for the training and validation of the network in 10-fold cross-validation. The atlas technique was also validated on this set. Inference dataset B consists of the remaining 26,753 quality-controlled subjects for whom no reference values were available. The network was applied to these subjects, but without reference measurements (and consequently no ground truth values) this dataset could not be used as a true independent test set. Instead, the atlas was applied to extract baseline measurements for comparison, but was too slow to process them all within the given time. Therefore, comparison dataset C was formed as a random subset of 1,000 subjects from dataset B and used to compare the network and atlas.

Image Formatting The stations of the neck-to-knee body MRI were fused and resampled to a common spatial resolution of 2.23mm $\times$ 2.23mm $\times$ 3mm, with $370 \times 224 \times 174$ voxels. Next, water and fat fraction images were calculated by voxel-wise division of the water or fat signal intensity by the sum of both signals. This sum image was also used to generate body masks by applying a threshold, calculated as the mean of Otsu filter thresholds for all coronal slices of the summed signal for a given subject.

For the neural network, a highly compressed two-dimensional format with coronal and sagittal fat fraction slices was extracted, as seen in Fig. 1. Based on the body mask, the coronal slice was extracted at center of mass and the sagittal slice at a quarter of mass, typically locating the latter along the center of the right thigh. Both slices were cropped to exclude the bed and the bottom half of the body and then concatenated. The resulting image of 376×176 pixels was then compressed to an 8bit format, with fat fractions ranging from 0 to 50%. No body mask was applied to this format. While there is no guarantee that this strategy captures actual liver tissue it operated robustly and required only about 5 seconds per subject with a GPU implementation.

Neural Network Configuration The convolutional neural network is based on a ResNet50 [5] modified for regression of a single value. Most hyperparameters of the generalized regression framework were retained [9], with batch size 32, online-augmentation by translations and weights pretrained on ImageNet. Accordingly, a mean squared error loss was optimized with Adam and a base learning rate of 0.0001 was reduced by factor ten in the last 1,000 iterations of 6,000 iterations in total. Each training sample consists of the two-dimensional image format as extracted from the neck-to-knee body MRI as input and the reference liver fat measurement of the same subject as ground truth value. Training on one split required about 25 minutes whereas prediction was almost instantaneous. All experiments ran on a Nvidia GTX 1080 Ti 11GB GPU, a 12-core Intel Xeon W-2133, 3.60 GHz CPU and 32GB RAM in PyTorch.

Fig. 1. Input format for the network as extracted from the neck-to-knee body MRI volumes. The upper halves of specific coronal and sagittal fat fraction slices are combined in a two-dimensional 8bit image of 376×176 pixels, encoding fractions of 0-50%. For the shown subject, the reference method lists a liver fat percentage of 15%.

Multi-Atlas Baseline A simple multi-atlas segmentation pipeline was implemented on the neck-to-knee body MRI, performing a median readout of minimal liver segmentations. Three subjects (one female, two male) with high observed variance in liver shape were manually segmented to serve as templates, using the water fraction and signal images to outline the liver and exclude vessels and adjacent tissue. When applying the atlas, these templates were transformed to each target with a graph-cut based deformable registration technique [3] on the full paired water and fat fraction volumes, after applying the body mask to remove background noise. A resolution pyramid of six levels was used together with a GPU-implementation of the normalized cross-correlation. The obtained deformation field for each template was applied to the corresponding liver segmentation. The thus aligned binary segmentations were then multiplied and subsequently eroded with a spherical kernel of seven voxels in diameter. The median fat fraction value of the remaining selected voxels was then returned. PDFF values of the neck-to-knee body MRI may not directly correspond to those reconstructed by the reference method [8]. As a final step, the raw atlas output was therefore fit to the reference values on the validation dataset A by linear regression. When applied to other datasets, the same parameters were applied.

Evaluation The network was trained to emulate the reference method by regression. The success of this training was therefore evaluated by quantifying the quality of fit with the coefficient of determination R^2 and reporting the mean absolute error (MAE) and the 95% limits of agreement (LoA). Furthermore, Pearson's coefficient of correlation r is used to examine the randomness of errors by network and atlas relative to the reference. The measured values can also be

thresholded to identify subjects above the NAFLD risk level of 5.5 FF. The area under curve (AUC) of the receiver operating characteristic (ROC) curve is reported together with sensitivity and specificity for thresholding at this level. For some outliers, manual segmentations similar to the atlas templates were created. Their extent was more conservative, and their median FF values are reported after correction with the same linear regression parameters as applied for the atlas.

3   Results and Discussion

Results are shown in Fig. 2 and Table 3. The network outperformed the atlas on validation dataset A but retained the same pattern of agreement with it on comparison dataset C, indicating robust generalization to those subjects lacking reference values. On dataset A, the network inferred outliers of up to 23 FF, with 16 subjects reaching errors above 5 FF. However, manual segmentation of the top ten outliers yielded measurements that agree for LoA of (-0.7 to 1.2 FF) to the atlas and (-5.4 to 7.3 FF) with the network, but only (-23.5 to 8.4 FF) with the reference. This substantial disparity shows a mismatch between the reference and liver fat as observed in the neck-to-knee body MRI in these subjects that can not be explained by the different imaging protocol alone, but might instead be the result of spurious outliers in the reference method. The errors of atlas and network relative to the reference on dataset A are highly correlated ($r = 0.715$ with LoA -1.9 to 2.0).

The agreement between network and reference with LoA (-2.2 to 2.3 FF) is well within the LoA of (-4.0 to 3.4 FF) reported between the gold standard modalities of MRI and MRS [18]. Even within MRI, variability with LoA of (-2.2 to 1.8 FF) for comparable protocols and (-2.5 to 1.6) between different imaging sites has to be expected [6]. The atlas was about fifty times slower than the network and affected by systematic differences between the two- and three-point Dixon techniques [8]. It was only evaluated after correction with the parameters from linear regression ($0.9x - 0.8$ FF), which failed to resolve a nonlinear structure in the range of 1-4 FF, however. Perhaps due to the outliers, the corrected values still tend to overestimate low and underestimate high values. In

Table 1. Method comparison on datasets A (first three rows) and C (bottom row). The three final columns assume thresholding at 5.5 FF, with the first named method as ground truth. MAE: mean absolute error, LoA: 95% limits of agreement, ROC-AUC: area under receiver operating characteristic curve, Sens: sensitivity, Spec: specificity.

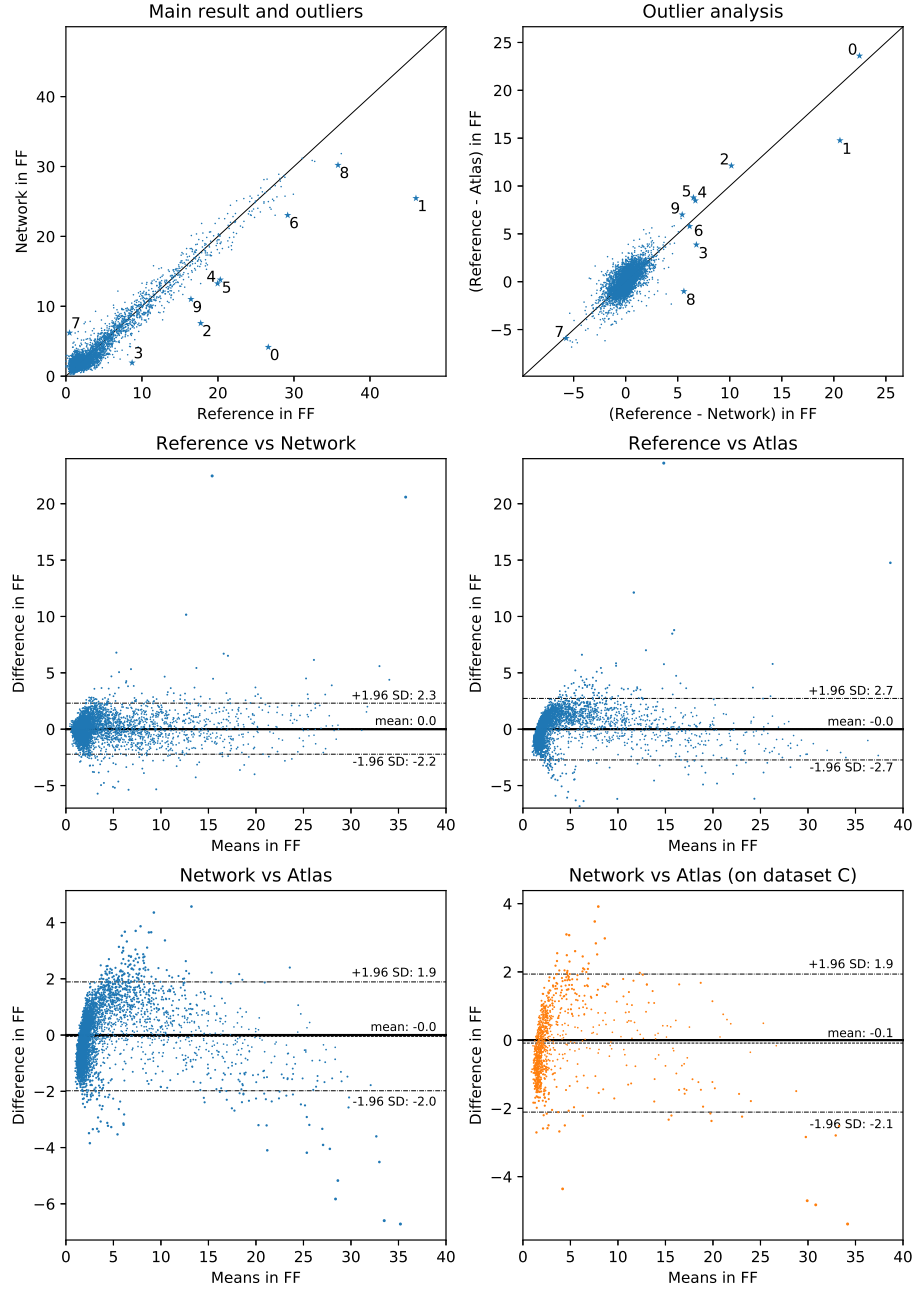|			MAE	R^2	LoA	ROC-AUC	Sens	Spec
Reference	vs	Network	0.77	0.940	(-2.22 to 2.31)	0.992	89.3	98.2
Reference	vs	Atlas	1.03	0.912	(-2.73 to 2.73)	0.991	78.0	99.2
Atlas	vs	Network	0.76	0.952	(-1.89 to 1.98)	0.995	97.4	95.8
Atlas (C)	vs	Network (C)	0.80	0.950	(-1.94 to 2.11)	0.991	93.6	95.6

Fig. 2. Results on datasets A (blue, N=4,418) and C (red, N=1,000). Differences were calculated by subtracting the second named method from the first. Dashed lines denote 95% limits of agreement. Both network and atlas independently contradict the reference in the annotated top ten network prediction outliers, indicating implausible reference values. Network inference on dataset C retains its agreement with the atlas as observed on dataset A, showing no signs of deteriorated performance in generalization.

contrast, the network learned to rectify values in the lower ranges and emulated the reference method better, possibly with information from additional image features. With LoA of (-2.7 to 2.7 FF), the atlas still surpasses the the generalized regression framework, which used mean intensity projections of water and fat signals for LoA of (-4.0 to 4.2 FF) on the same UK Biobank subjects [9].

Several limitations apply to the presented results. The comparison on dataset C indicates robust generalization to UK Biobank subjects from any of the given centers with the chosen protocol. However, without an independent test set no conclusions can be drawn for other demographic groups, scanning devices or imaging protocols, which would likely at least require retraining of the network. The slice selection strategy is also not guaranteed to be optimal or even certain to capture any liver tissue, and inherently limited by encoding only values from 0-50 FF in 8bit. Conceptually, the potential of deep learning on the neck-to-knee body MRI is still not fully leveraged, as design choices regarding the compression of the volumes into two-dimensional representations still strongly affects the results, reminiscent of hand-crafted features. The proposed regression technique also generates no output segmentations, so that individual predictions can not be easily explained or corrected. In contrast, neural networks for segmentation have been proposed for automation of parts of the reference method [7] and could provide these, reducing any manual work to the initial creation of representative ground truth segmentations for training only.

However, apart from the underlying design decisions, the presented approach eliminates the need for manual intervention entirely. No human guidance, model-based assumptions or representative ground truth segmentations are required, and inference for thousands of subjects can be performed within a day. Future work will consist in expanding the training and evaluation samples, once available, which is likely to further improve performance [10]. Likewise, the planned repeat scans of up to 70,000 subjects could potentially be processed without any further changes to the presented system, enabling convenient longitudinal samples of liver fat. The reference method also provides a liver inflammation factor and iron content, and while there is no indication that these could be inferred from the neck-to-knee body MRI in the same way, the presented approach could also be evaluated on the dedicated liver MRI slices directly. Furthermore, the inferred values of dataset B could be examined more thoroughly, using known correlations to the metadata and by expanding the baseline measurements.

It is worth noting that after the completion of these experiments an alternative reference set of about 10,000 image-based liver fat measurements has been released by the UK Biobank as field 22436-2.0. A full analysis of these values is beyond the scope of this work, but preliminary results indicate that field 22436-2.0 may be of higher quality than field 22402-2.0 used here. Like the network and atlas, the new reference contradicts field 22402-2.0 on the top ten outliers found in this work. Early attempts to emulate this new reference with the proposed method yielded a closer fit. As the data collection is continuously progressing, future work will be able to examine the relationship between both references and the proposed technique in ever increasing sample sizes.

4   Conclusion

In conclusion, the proposed framework can emulate the reference measurements and infer similar liver fat values from the neck-to-knee body MRI of the UK Biobank. It outperforms the atlas and combines high speed with accurate and objective predictions, while eliminating any need for manual intervention or guidance and leaving no room to subjective variability. The inferred liver fat measurements are readily available at large scale for distribution as return data by the UK Biobank. They could be used to identify potentially erroneous outliers as observed in the reference method, but also as an approximation in medical research, enabling larger sample sizes until all subjects have been evaluated with more established gold standard techniques.

References

1. Browning, J.D., Szczepaniak, L.S., Dobbins, R., Nuremberg, P., Horton, J.D., Cohen, J.C., Grundy, S.M., Hobbs, H.H.: Prevalence of hepatic steatosis in an urban population in the united states: impact of ethnicity. Hepatology **40**(6), 1387–1395 (2004)
2. Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage **163**, 115–124 (Dec 2017). https://doi.org/10.1016/j.neuroimage.2017.07.059, `https://linkinghub.elsevier.com/retrieve/pii/S1053811917306407`
3. Ekström, S., Malmberg, F., Ahlström, H., Kullberg, J., Strand, R.: Fast Graph-Cut Based Optimization for Practical Dense Deformable Registration of Volume Images. arXiv:1810.08427 [cs] (Oct 2018), `http://arxiv.org/abs/1810.08427`, arXiv: 1810.08427
4. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., Kitamura, F.C., Thodberg, H.H., Chen, L., Shih, G., Andriole, K., Kohli, M.D., Erickson, B.J., Flanders, A.E.: The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology **290**(2), 498–503 (Feb 2019). https://doi.org/10.1148/radiol.2018180736, `http://pubs.rsna.org/doi/10.1148/radiol.2018180736`
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (Jun 2016). https://doi.org/10.1109/CVPR.2016.90
6. Hernando, D., Sharma, S.D., Aliyari Ghasabeh, M., Alvis, B.D., Arora, S.S., Hamilton, G., Pan, L., Shaffer, J.M., Sofue, K., Szeverenyi, N.M., et al.: Multisite, multivendor validation of the accuracy and reproducibility of proton-density fat-fraction quantification at 1.5 t and 3t using a fat–water phantom. Magnetic resonance in medicine **77**(4), 1516–1524 (2017)
7. Irving, B., Hutton, C., Dennis, A., Vikal, S., Mavar, M., Kelly, M., Brady, J.M.: Deep quantitative liver segmentation and vessel exclusion to assist in liver assessment. In: Annual Conference on Medical Image Understanding and Analysis. pp. 663–673. Springer (2017)
8. Kukuk, G.M., Hittatiya, K., Sprinkart, A.M., Eggers, H., Gieseke, J., Block, W., Moeller, P., Willinek, W.A., Spengler, U., Trebicka, J., et al.: Comparison between

modified dixon mri techniques, mr spectroscopic relaxometry, and different histologic quantification methods in the assessment of hepatic steatosis. European radiology **25**(10), 2869–2879 (2015)

9. Langner, T., Ahlström, H., Kullberg, J.: Large-scale biometry with interpretable neural network regression on uk biobank body mri. arXiv preprint arXiv:2002.06862 (2020)

10. Langner, T., Wikström, J., Bjerner, T., Ahlström, H., Kullberg, J.: Identifying morphological indicators of aging with neural networks on large-scale whole-body MRI. IEEE Transactions on Medical Imaging pp. 1–1 (2019). https://doi.org/10.1109/tmi.2019.2950092, `https://doi.org/10.1109/tmi.2019.2950092`

11. Pagadala, M.R., McCullough, A.J.: The relevance of liver histology to predicting clinically meaningful outcomes in nonalcoholic steatohepatitis. Clinics in liver disease **16**(3), 487–504 (2012)

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

13. Reeder, S.B., Cruite, I., Hamilton, G., Sirlin, C.B.: Quantitative assessment of liver fat with magnetic resonance imaging and spectroscopy. Journal of magnetic resonance imaging **34**(4), 729–749 (2011)

14. 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.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine **12**(3), e1001779 (Mar 2015). https://doi.org/10.1371/journal.pmed.1001779, `https://dx.plos.org/10.1371/journal.pmed.1001779`

15. West, J., Dahlqvist Leinhard, O., Romu, T., Collins, R., Garratt, S., Bell, J.D., Borga, M., Thomas, L.: Feasibility of MR-Based Body Composition Analysis in Large Scale Population Studies. PLoS ONE **11**(9) (Sep 2016). https://doi.org/10.1371/journal.pone.0163332, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5035023/`

16. 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.: Characterisation of liver fat in the uk biobank cohort. PloS one **12**(2), e0172921 (2017)

17. Xue, W., Islam, A., Bhaduri, M., Li, S.: Direct multitype cardiac indices estimation via joint representation and regression learning. IEEE transactions on medical imaging **36**(10), 2057–2067 (2017)

18. Yokoo, T., Serai, S.D., Pirasteh, A., Bashir, M.R., Hamilton, G., Hernando, D., Hu, H.H., Hetterich, H., Kühn, J.P., Kukuk, G.M., et al.: Linearity, bias, and precision of hepatic proton density fat fraction measurements by using mr imaging: a meta-analysis. Radiology **286**(2), 486–498 (2018)

19. Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., Wymer, M.: Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology **64**(1), 73–84 (2016)