Supervoxel-based image-to-biomarker conversions - An initial study on morphological age prediction from whole-body MRI and its clinical relevance in the UK Biobank

Johan Öfverstedt[1], Alexander Blomlöf[1], Yasemin Utkueri[1], Rama K. Guggilla[1], Elin Lundström[1], Håkan Ahlström[1,3], Joel Kullberg[1,2,3]

1 Department of Surgical Sciences, Uppsala University, Sweden.
2 Department of Surgical Sciences, SciLifeLab, Uppsala University, Sweden.
3 Antaros Medical, Mölndal, Sweden
Corresponding author: johan.ofverstedt@uu.se

Abstract:

Biological aging remains a central focus of research, from the scale of sub-cellular processes to whole-organism tissue morphology and function. In this work, we developed a novel and quantitatively interpretable method for the prediction of variables, such as age, from tomographic medical images. The method uses supervoxels (whose granularity is selected by the user), standardized through inter-subject image registration, and tissue-specific feature extractions from each supervoxel to convert all image data collected into a set of well-defined imaging biomarkers. We applied the method to age prediction, using linear modelling and whole-body water-fat MRI data of 38,235 subjects (age span = 45 - 82 years) in the multicentre UK Biobank study, resulting in predictions representing morphological age (MA). We observed state-of-the-art whole-body age prediction performance on a held-out test set with a mean absolute error of 1.951/2.057 years, and R^2 of 0.884/0.892 for females/males, respectively. The method was observed to outperform both previously reported CNN-based results from the UK Biobank, and predictions from explicit biomarkers, from a multi-organ/tissue segmentation approach, in a direct comparison. The image-based interpretability of the model allowed for detailed and tissue-specific analysis of the utilized body-wide associations with age. Important regions for the predictions included volumes of the aorta, regional muscle, bone marrow, and adipose tissue depots, and lean tissue fat content. The predicted MAs were of clinical relevance as they were significantly related to both type 2 diabetes and all-cause mortality. A key finding was an accuracy/utility trade-off where the more parsimonious models showed lower CA predictive performance but higher clinical relevance and interpretability.
The proposed method facilitates automated image-to-biomarker conversions and predictions based on subsets of anatomies, tissues, and image features, for potential application in numerous future medical studies. Code is shared at github.com/Radiological-Image-Analysis-Group-UU/ukbb_age_prediction.

Introduction

Biological age (BA) is a major risk factor for a wide spectrum of human diseases, including cancer, diabetes, and cardiovascular disorders (1). While chronological age (CA) represents the time since birth, BA reflects an individual's physiological state, which is highly associated with chronological age but also affected by, for example, diseases, genetic factors, and lifestyle factors (26, 28). BA can be estimated through various approaches, including imaging (14), functional assessments, metabolomics, proteomics, and DNA methylation. BA has been shown to be a stronger predictor of chronic health outcomes than CA (2, 30).

Identifying MR imaging biomarkers that capture aspects of biological aging provides an opportunity for anatomically localized biomarker discovery of importance to multiple diseases. Geroprotective interventions also require proxy methods to evaluate therapy efficacy (9, 21).

The UK Biobank (UKBB) is an ongoing longitudinal study initiated in 2006 (3). Participants across the UK have contributed extensive data, including genetics, biological samples, and Magnetic Resonance (MR) images. Leveraging MR images for the non-invasive acquisition of disease-associated imaging biomarkers is vital for advancing numerous targeted prevention strategies, developing effective therapies, and improving patient outcomes (29). UKBB MR images have previously been utilized for age predictions (for varying definitions of age), where different body regions have been used as input: neck-to-knee MR images (4, 5), abdominal MR images (7, 8), and brain MR images (6, 7, 24).

Tomographic imaging techniques collect vast amounts of information about the investigated subjects and the optimal use of this detailed information poses a significant challenge. In clinical practice, visual interpretation or simple measurements based on delineation of regions of interest are commonly used. Such delineations may be task-specific or done by more general-purpose segmentation models, such as the recent VIBESegmentator (17), which delineates 71 defined regions corresponding to various organs and tissues in whole-body MR images. Supervoxels are partitionings of 3D images based on spatial proximity and similarity of voxel intensity, providing compressed image representations. Typically, no semantic information about regions or structures of interest is required for creating the supervoxel representations, which makes them versatile and powerful. However, this lack of semantic grounding makes it challenging to relate supervoxel-based measurements between subjects. Inter-subject image registration is a process for spatial standardization of images, of which one use case is to standardize supervoxels to be mapped to well-defined anatomical regions between subjects. However, spatial standardization, which is routinely performed in some subfields, such as neuroscience, poses a challenge for whole-body images due to greater inter-subject anatomical heterogeneity compared to brain scans.

Convolutional neural networks (CNNs) have great potential in mining image data in a data-driven manner (23). Challenges, however, include an extensive need for training data, model overfitting, and shortcut learning (22, 29). Even though explainability approaches have

been proposed, such as post-hoc saliency analysis, the interpretation of these models remains a significant challenge (25, 27).

The aims of this study were:

- To develop and evaluate a novel method based on tissue-specific standardized supervoxel prediction (TS-SSP) and apply it for age prediction from whole-body water-fat MR images, here referred to as morphological age (MA).
- To compare the performance of the standardized supervoxel-based features against features extracted using a state-of-the-art tissue segmentation method.
- To investigate the importance of tissue-specific feature extraction.
- To interpret the prediction models and analyze how different regions and features are associated with MA.
- Evaluate the relevance of the predicted MAs for diabetes and all-cause mortality.

Materials and methods

An illustration of the main steps of the method is presented in Fig. 1.

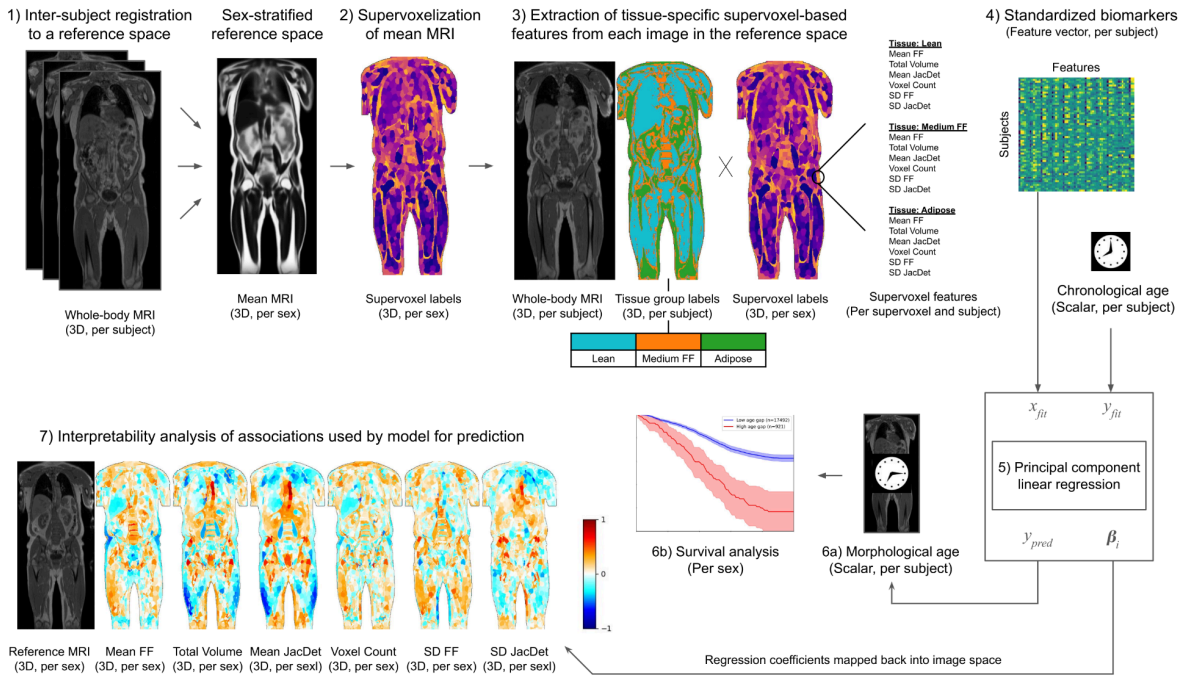

Figure 1: An overview of the proposed tissue-specific standardized supervoxel prediction *(TS-SPP) method. The following steps are performed for females and males separately. 1) Whole-body Dixon MR images (with a water signal and fat signal image component from which fat fraction images are calculated) are registered and transformed into a common reference image space. 2) Supervoxels are constructed in the common image space. The supervoxels are created using a mean image calculated from all registered images. 3) Three tissue groups are defined using thresholding of fat fraction values in each original image, partitioning the voxels into groups of voxels with low, medium, and high fat content. 4) Interpretable features related to the intensities (fat fractions) and original tissue volumes (estimated from the registration through Jacobian determinants, i.e., JacDet, of the deformation fields) are calculated from within each supervoxel per tissue group (separately for the three tissue groups). The extracted data points are referred to as "supervoxel features" and can be described as extracted imaging biomarkers, all with a defined and interpretable meaning shared between all subjects and images. 5) A regression model is trained to predict a subject's chronological age from these supervoxel features. 6a) The performance of the predicted morphological ages (MAs) is evaluated, and 6b) the association of the predicted MAs with health-related variables is evaluated. 7) Cohort-level interpretability analysis is conducted by mapping the model's used association of each feature, tissue group (using the reference subject's tissue group assignment for each voxel), and supervoxel combination, in the common reference image space for a sub-supervoxel visualization of the models.*

A total of 38,235 visually inspected and approved whole-body, neck-to-knee MR images from the UKBB (1) were used in this study, consisting of 19,796 female subjects and 18,439 male subjects. The mean and standard deviation of age were 63.6 ± 7.4 years for females and 64.9 ± 7.6 years for males. The sex-stratified subjects were divided into ten folds per sex at random to enable cross-validation (CV), with one fold per sex put aside as a held-out test set. The test set consisted of 1,980 females and 1,844 males. The research was approved by the national ethical committee (Dnr: 2019-03073).

The MR data were acquired using dual-echo Dixon imaging on Siemens Aera 1.5T scanners across four UK imaging centers, providing separate but inherently coregistered water signal and fat signal images. The neck-to-knee coverage was achieved using imaging across six stations, each capturing a specific region of the body. The subvolumes for both the water and fat signals from each station were stitched into a common space, resulting in 3D volumes of size 362 x 174 x 224 voxels. Fat fraction (FF) volumes were calculated by element-wise division of the fat volumes by the sum of the water and fat volumes.

Inter-subject deformable image registration to a reference image was performed to map all the images into a common space (10, 11), as seen in Fig. 1. One male and one female subject were designated reference subjects, selected based on being representative in terms of body size and body shape (both with CA≈64 years). The registration was performed on water and fat fraction volumes, along with subcutaneous fat and muscle masks generated by VIBESegmentator (17), applied on the sum of the water and fat images. For each deformation field, a deformed FF image was computed along with the Jacobian determinant (JacDet) field representing the local estimated reference voxel-wise expansion or contraction. More details are presented in Supplementary A.

A sex-stratified mean FF volume was created and used as input to generate a supervoxelization (to reduce bias, compared to using a single image) with the Simple Linear Iterative Clustering (SLIC) algorithm (13), which generates a configurable number of supervoxels (more details are found in Supplementary B). Figure 2 shows the mean FF volumes and supervoxelizations for various choices of the number of supervoxels.

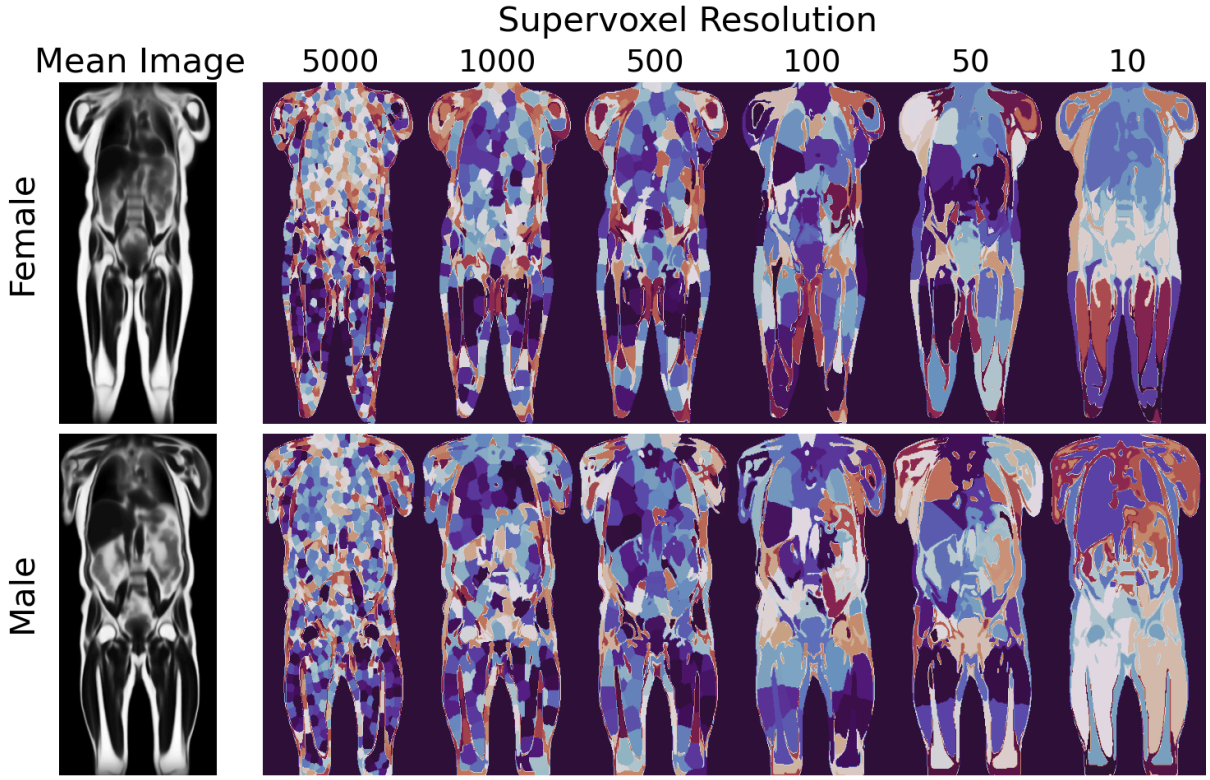

Figure 2. Visualization of the mean image for each sex on the left, and different supervoxelizations on the right corresponding to varying numbers of supervoxels. A higher number of supervoxels provides a more detailed partitioning of the body, a larger number of features, each computed from a smaller number of voxels.

For each subject, we extracted features from the voxels within each standardized supervoxel (20). To alleviate partial volume effects resulting from the presence of multiple tissue types (e.g., both adipose tissue and lean tissue) within a single supervoxel, we propose extracting features specific to user-defined tissue groups. The tissue-specific features were selected based on CV

performance. These were: mean FF, total volume (i.e. mean JacDet*voxel count), mean JacDet, reference space voxel count, standard deviation (SD) FF and SD JacDet. See detailed definitions in Supplementary C.

We defined three tissue groups based on FF, here denoted “low FF” (lean tissue), “medium FF”, and “high FF” (adipose tissue), for a method referred to as the *3-tissue group configuration*. These thresholds were selected empirically based on the histogram of FF values. Denoting x a voxel, $I(x)$ the FF of voxel x, and $T_(x)$ tissue group weights, these 3 tissue groups were defined as binary categorizations:
$T_{\mathbf{low}}(x) = 1$ if $I(x) \leq 20\%$, 0 otherwise, $T_{\mathbf{high}}(x) = 1$ if $I(x) > 80\%$, 0 otherwise, and $T_{\mathbf{medium}}(x) = 1 - T_{\mathbf{low}}(x) - T_{\mathbf{high}}(x)$. For comparison, we also consider a version of the method where all the voxels were assigned to a single tissue group ($T_{\mathbf{all}}(x)=1$), referred to as the *1-tissue group configuration*, equivalent to feature extraction without tissue groups.

We applied Z-score normalization with clipping (below -3 SD, above 3 SD) to the features (20). Principal component analysis (PCA) dimensionality reduction with 3000 components was applied to the z-scores to combine correlated features into components (20), and to reduce the risk of shortcut learning (22). Additionally, we selected a subset of 10 out of 3000 PCA components, maximizing the univariate absolute correlation with CA. These operations were performed fold-wise based on training set statistics. Linear models, selected for both simplicity and interpretability, were fitted to predict the CA from the PCA components using 9-fold CV. To visualize the used associations and interpret the predictions, the regression coefficients were mapped to each supervoxel/tissue/feature combination z-score, normalized through division by the maximum absolute value, and plotted in the reference space.

The code for the method is shared as open source (19).

Evaluation and statistics

The performance of the prediction models was evaluated on both the validation data and on the held-out test set, with the CA as reference. The measures used to quantify the prediction errors were 1) the mean absolute error (MAE), which decreases with improved performance, and 2) the coefficient of determination R^2, which increases with improved performance. Ablation studies were conducted to examine how age predictions were influenced by training set size, number of supervoxels, the impact of the 3-tissue group configuration compared to the 1-tissue group configuration, as well as tissue-specific experiments where age was predicted from features extracted from a 1-tissue group and from a single tissue group and feature combination.

To evaluate the advantage of the standardized supervoxel features in comparison to alternative available segmentation approaches, we also considered a recently published method, VIBESegmentator, which segments whole-body MRI scans into 71 regions comprising organs and tissues. From these segmentations and tissue-group masks, we extracted features analogously to the supervoxel-based features.

We performed statistical tests to compare the absolute errors of the 3-tissue group configuration (TS-SSP) compared to the 1-tissue group configuration, both using supervoxels, as well as the 3-tissue group supervoxel method compared to the 3-tissue group VIBESegmentator method. 36 tests (9 per sex and method) were performed; therefore, Bonferroni correction with a factor of 36 was applied to adjust the p-values.

To evaluate the individual importance of each tissue and feature, tissue- and feature-wise predictions were performed.

Interpretability analysis was performed on both the 10-component and 3000-component models by reading of the 3D interpretability maps (30) by a radiologist with 40 years of experience.

In addition to the evaluation of the age predictions, we investigated the relevance of the predicted ages to the prevalence of type-2 diabetes (T2DM) and all-cause mortality. After bias correction (described in detail in Supplementary G), which removes systematic underestimation of high ages and overestimation of low ages (24), we computed an age gap as the difference between each bias-corrected predicted age and the subject's CA (age gap = MA-CA).

We plotted a moving average of the age gaps for the subjects with T2DM as a function of CA. Bootstrapping (with 1000 repetitions) was applied to the moving averages to obtain 95% confidence intervals (CI) of the average age gaps. We also performed logistic regression with either CA or MA as input, and comparing the AUC values were compared with DeLong's test with Bonferroni correction factor of 2.

The survival analysis consisted of Kaplan-Meier plots with two groups: 1) subjects with an age gap larger than the top 10th percentile, and 2) remaining subjects. Groupwise bootstrapping (with 10000 repetitions) was applied to the two groups to obtain a 95% CI. Statistical testing with

the log-rank test was performed with a Bonferroni correction factor of 2. For comparison of the CA and MA as predictors of mortality, we also performed logistic regression to predict 5-year all-cause mortality with either CA and MA as independent variables, and the resulting area under the curve (AUC) values were compared with DeLong's test with Bonferroni correction factor of 2.

Results

The performance of the proposed method in comparison to the reference CAs is presented in Table 1, where the 3-tissue group configuration significantly outperformed the 1-tissue group configuration in terms of MAE and R^2, both on the cross-validation experiments and on the test set, except for fold 6 in males (p-values are shown in Supplementary H).

Table 1: The performance of the two primary models (with 3000 PCA components and 5000 supervoxels) in cross-validation (CV) experiments and on the test set, which consists of 1,980 samples in the female cohort and 1,844 samples in the male cohort. Performance was measured using R^2 and MAE, with CA as reference. Each cell presents the mean results of nine models on the respective partition, fitted using cross-validation, with the standard deviation across models in parentheses.

Tissue groups	*Partition*	*R^2 Female*	*R^2 Male*	*MAE (years) Female*	*MAE (years) Male*
1-tissue group	CV	0.868 (0.004)	0.871 (0.005)	2.131 (0.026)	2.179 (0.024)
3-tissue group [TS-SSP]	CV	**0.882 (0.004)**	**0.886 (0.003)**	**2.006 (0.028)**	**2.043 (0.022)**
1-tissue group	Test set	0.869 (0.001)	0.879 (0.001)	2.074 (0.006)	2.166 (0.011)
3-tissue group [TS-SSP]	Test set	**0.884 (0.001)**	**0.892 (0.001)**	**1.951 (0.009)**	**2.057 (0.014)**

The cohort-level interpretability analysis for the 10-component (exhibiting strong association to disease and survival) and 3000-component (exhibiting strong performance on CA prediction) 3-tissue group configuration of the method is visualized in Fig. 3 and Fig. 4. See also movies of the full 3D results in (16).

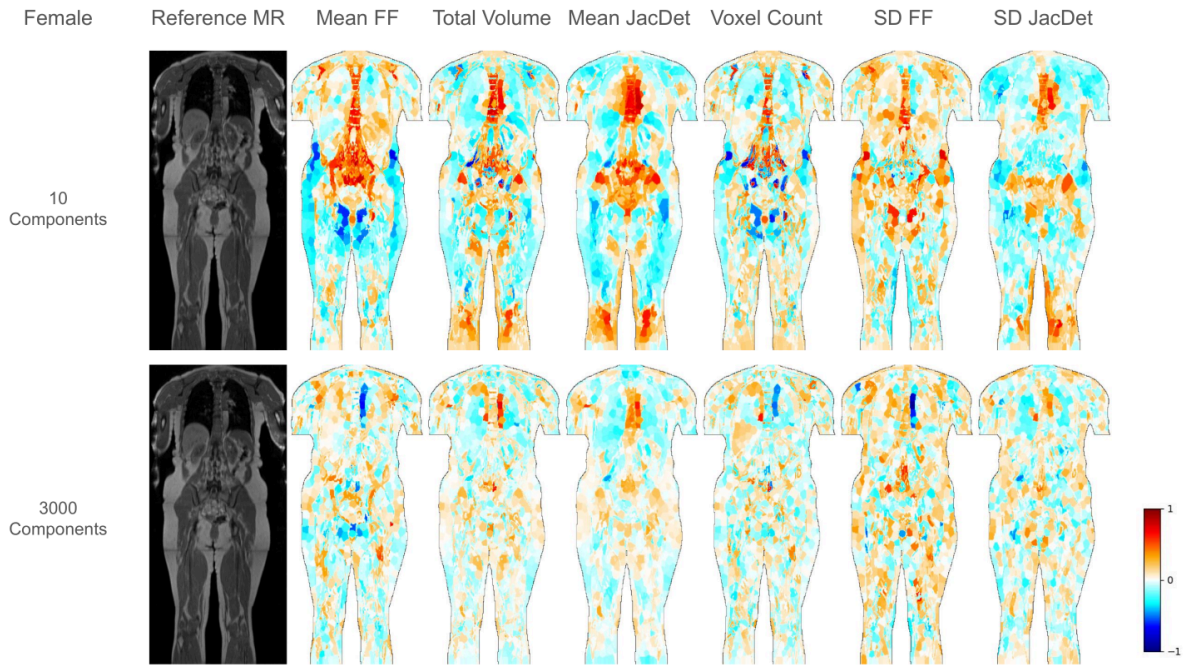

Figure 3: Cohort-level interpretability analysis for the female subjects. The leftmost column of each subfigure shows a slice from the reference volume, and the six subsequent images on each row show the association used by the models. The associations are normalized and expressed in standard deviations of each supervoxel feature. The tissue group information within each supervoxel in the reference subject is used to visualize the results on a sub-supervoxel level. The 10-component models exhibit a strong association with disease and survival (shown in Fig. 6 and Fig. 7), and the 3000-component models exhibit strong performance on CA prediction (shown in Tab. 1).

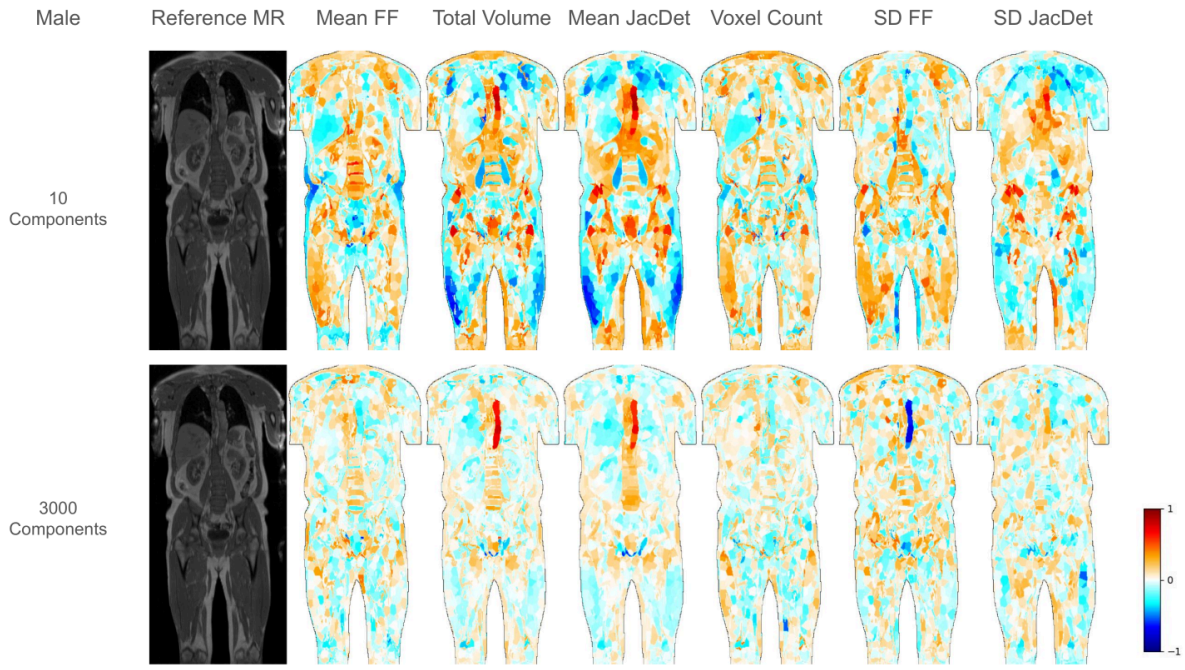

Figure 4: Cohort-level interpretability analysis for the male subjects. The leftmost column of each subfigure shows a slice from the reference volume, and the six subsequent images on each row show the association used by the models. The associations are normalized and expressed in standard deviations of each supervoxel feature. The tissue group information within each supervoxel in the reference subject is used to visualize the results on a sub-supervoxel level. The 10-component models exhibit a strong association with disease and survival (shown in Fig. 6 and Fig. 7), and the 3000-component models exhibit strong performance on CA prediction (shown in Tab. 1).

For the 10-component models, we observed a positive model association between MA and the volume of the aorta, prostate (in males), visceral adipose tissue (VAT), vertebrae, bone marrow, pulmonary artery (at the bifurcation), gluteus maximus (more distinctly in females), and lower versal abdominal SAT (more distinctly in females). We observed a negative model association between age and the volume of the liver, spleen, lungs, pancreas (more distinctly in females), kidneys, heart ventricles, upper versal abdominal SAT, muscles of the neck, back, and in the thighs, pulmonary artery (at the origin), femurs (predominately negative).

We observed a positive model association between age and mean FF of the kidneys, spleen, lower ventral abdominal SAT, pancreas, vertebrae, bone marrow, humerus, femur (weak), most muscles (distinctly in the back muscles). We observed a negative model association between age and the mean FF of the liver, VAT, dorsal and side abdominal SAT and (predominately negative) leg SAT.

Comparing the 10- and 3000-component models, the 10-component models exhibit anatomically coarse associations, and the 3000-component models exhibit more anatomically

detailed associations, with more intra-tissue and intra-organ variations. Several regions exhibit a change of sign of the association between the configurations. The volume of the aorta was a key feature in both configurations for both sexes.

The impact of the number of supervoxels and dataset size on age prediction performance is shown in Figure 5, and in Supplementary I, respectively, where more supervoxels and more data led to improved performance.

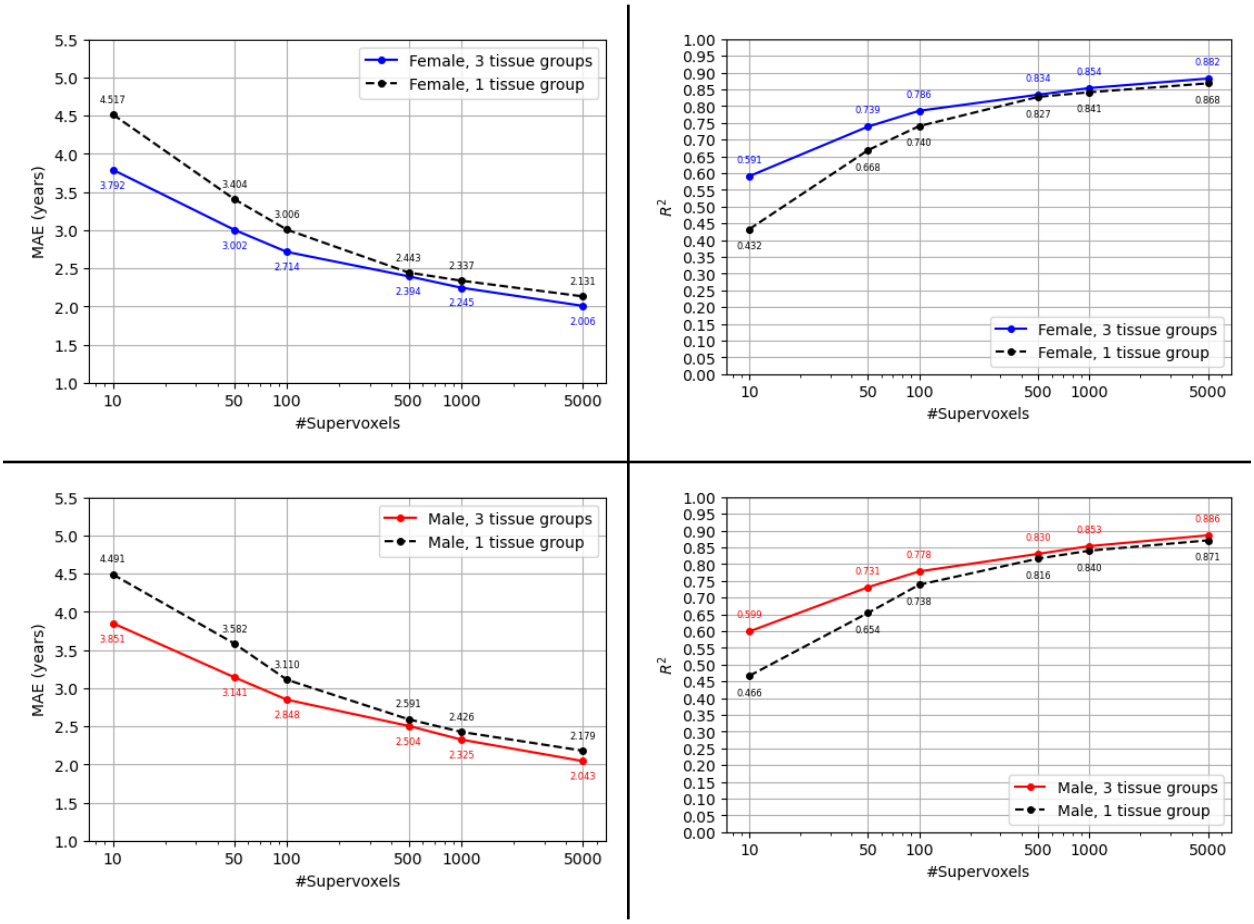

Figure 5: The effect of different numbers of supervoxels on MAE and R^2 in the CV evaluation of age prediction (evaluated on the CV sets), with a comparison of the 1-tissue group configuration with the 3-tissue group configuration. A lower MAE and a higher R^2 indicate better performance. The y-axis is scaled linearly, while the x-axis is scaled logarithmically.

The performance comparison between supervoxel features and VIBESegmentator-based features is shown in Table 2, where the supervoxel-based predictions outperform the segmentation-based prediction. We observed that the 3-tissue-group feature extraction also aided the performance of the VIBESegmentator-based age predictions. The p-values are shown in Supplementary H.

Table 2: Performance comparison between the main proposed models, which perform predictions using features extracted from 5000 supervoxels, and a model that performs predictions using features extracted from masks produced by VIBESegmentator. The performance was measured using R^2 and MAE. Each cell presents the mean results of nine cross-validation models on the test set, with the standard deviation across models in parentheses.

Tissue groups	*Regions*	*R^2 Female*	*R^2 Male*	*MAE (years) Female*	*MAE (years) Male*
1-tissue group	Standardized supervoxels	0.869 (0.001)	0.879 (0.001)	2.074 (0.006)	2.166 (0.011)
1-tissue group	VIBESegmentator	0.676 (0.001)	0.720 (0.001)	3.302 (0.004)	3.257 (0.003)
3-tissue groups [TS-SSP]	Standardized supervoxels	**0.884 (0.001)**	**0.892 (0.001)**	**1.951 (0.009)**	**2.057 (0.014)**
3-tissue groups	VIBESegmentator	0.712 (0.001)	0.764 (0.001)	3.115 (0.008)	3.035 (0.009)

35 tests out of the 36 were statistically significant in the comparisons between 3-tissue groups and 1-tissue group, and supervoxel-based features and VIBESegmentator-based features. The 3-tissue group configuration was superior to the 1-tissue group configuration except for male fold 6, and 3-tissue-group supervoxel-based features outperformed 3-tissue-group VIBESegmentator features.

Results from the tissue- and feature-wise age predictions are listed in the supplementary M. The numerically best performing predictions were found to be voxel count in medium FF tissue and the weakest was SD JacDet in adipose tissue.

The results of the association of the age gap and prevalence of T2DM are presented in Fig. 6, where we observed a substantial positive age gap in the subjects with T2DM compared to the control group. Prediction of T2DM was found to improve by using MA instead of CA. Logistic regression yielded an AUC delta of 5.5% for males (63.4% compared to 57.9%), p=0.0000, and 1.4% for females (63.0% compared to 61.6%), p=0.0475.

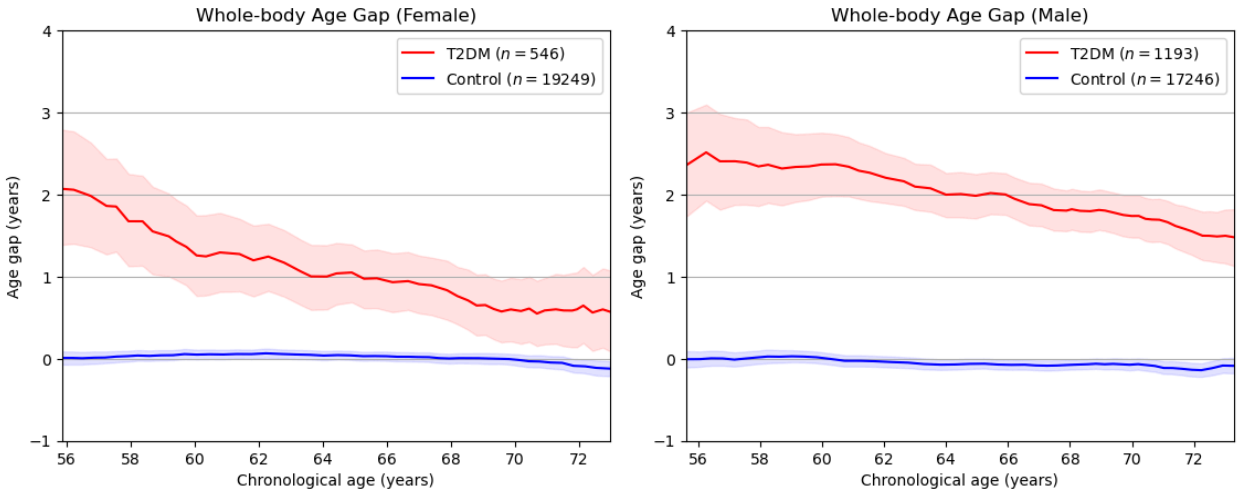

Figure 6: Association of type-2 diabetes (T2DM) and the predicted (and bias-corrected) age gap (age gap = MA - CA), as a function of CA, computed as a sliding window with average bandwidth ± 10 years. The age gap was corrected to a value close to zero for all ages in the control group. The T2DM group exhibits a substantially higher mean age gap for all CAs compared to the controls. CIs were generated by bootstrapping with 1000 samples.

The results of the survival analysis are presented as Kaplan-Meier plots in Fig. 7.

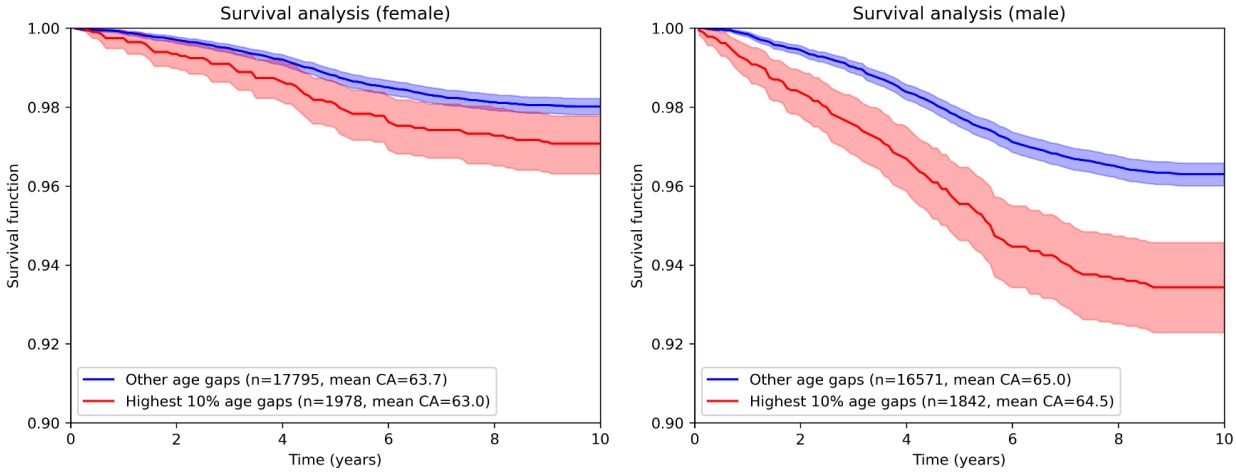

Figure 7: Survival analysis based on the predicted MAs. The highest 10% age gaps denote the group of subjects with an age gap larger than the top 10th percentile in each sub-cohort. The survival function models the number of subjects surviving beyond a given time point. A statistically significant relationship between age gap and survival is observed in both males and females, but is more pronounced in males. 10-year or longer follow-up time is available for all analyzed subjects. CIs were generated by bootstrapping with 1000 samples.

Subjects with the largest age gap (larger than the 10th percentile) have a significantly higher all-cause mortality compared to the rest of the cohort (adjusted p-value for males: $<2.86 * 10^{-9}$, adjusted p-value for females: < 0.0096).

The prediction of all-cause mortality from CA and MA is presented in Table 3.

Table 3: All-cause mortality prediction with logistic regression from CA and MA. The Bonferroni corrected P-values (with correction multiple 2) are presented along with the AUC values for CA and MA.

	Subjects	Events	CA AUC	MA AUC	P-value
Female	19773	252	0.698	0.705	0.479
Male	18413	456	0.719	0.733	0.005

We found a statistically significant difference between CA and MA as predictors of all-cause mortality for males (456 events), and no statistically significant difference for females (252 events).

Discussion

We developed an image-to-biomarker method for tissue-specific spatially standardized predictions (TS-SSP) using supervoxel-based interpretable features and applied it to the prediction of MA from neck-to-knee Dixon MR images in the UKBB, additionally evaluating its clinical relevance in terms of T2DM and all-cause mortality. The method achieved an MAE of 1.951 years on the female cohort (*n=1980*) and an MAE of 2.057 years on the male cohort (*n=1844*), evaluated in comparison to CA, also outperforming features extracted using the state-of-the art segmentation method VIBESegmentator. To the best of our knowledge, these results represent the highest R^2 and lowest MAE on age prediction achieved from neck-to-knee MR images. However, we note that, e.g., subset sizes, age ranges, and disease prevalence make direct comparisons to other published results challenging.

The two most similar studies to ours, Langner et al. (4) and Starck et al. (5), also used neck-to-knee MR images from the UKBB, without sex-stratification. Langner et al. utilized 23,000 images in a 10-fold cross-validation setup, and Starck et al. used a training set of 1,536 images. On their designated test sets, Langner et al. and Starck et al. achieved a mean absolute error (MAE) of 2.47 years and 2.57 years, respectively. Both studies conducted saliency analyses using post-hoc explainability methods, aggregating the saliency maps using image registration. The models by Langner et al. predominantly highlighted the aortic arch and the knees, while the model by Starck et al. emphasized the spine, the autochthonous back muscles, and the heart, along with its adjacent vessels. In another study, age was predicted from MR brain scans acquired of 15,000 healthy subjects, using a patch-based ensemble approach (6), yielding a MAE of 1.96 years and R^2 of 0.87. This performance is comparable to what is achieved in this study, even though we hypothesize that predicting the CA from healthy subjects is a less challenging task than when including subjects with, e.g., cardiometabolic diseases. In a study with chest CT scans as input, with ages between 55 and 75 years, an MAE of 1.84 years was achieved (lower MAE but narrower age range), also observing relevance of the predicted ages to outcomes in cancer patients (28).

In this study, we have fitted linear models to supervoxel-based standardized imaging biomarkers to predict clinically relevant MAs as well as generate 3D interpretability maps to analyze the model associations. The number of supervoxels can be adjusted with a single parameter, for a trade-off between the number of features and the spatial resolution, and tissue-specific supervoxel features can reduce the effect of mixing of different tissues. In the absence of reference MAs, we relied on a hypothesized capacity of the models to learn the patterns of normal aging from the CAs. We observed that for 3000 PCA components, the CA prediction exhibited very high performance. However, the resulting MAs did not exhibit a strong association with T2DM and mortality. Using a more parsimonious model with the 10 strongest components, the performance on CA prediction was inferior, but the association with T2DM and mortality was stronger and statistically significant, and the resulting interpretability maps were found to be more interpretable. We conclude that the models achieving high CA predictive performance may lose information about pathology and morphological aging, and thereby be less clinically relevant. On the other hand, the models with inferior CA predictive performance

can be useful for estimating MA and its association with diseases and all-cause mortality. Future studies need to consider this balance between prediction accuracy and interpretability. In addition, age prediction studies need to consider this well known challenge of tracking biological age models using CA.

This work has several limitations. Inter-subject registration is challenging due to differences in anatomy and appearance, and the resulting errors and ambiguities may present as feature noise. Bias can be introduced in the selection of the reference scan, although some bias is mitigated by deriving the supervoxels from the mean image. Partial volume effects may still exist within the defined tissue groups. Analysis of the association between, in particular, mean FF and age is complicated by the use of FF to define the tissue groups. A subset of the images contains stitching artifacts and undetected water-fat swaps (i.e., misidentification of water-based tissue as fat-based tissue and vice versa), which may influence both the image registration and feature extraction. For females, both the number of deceased subjects and the number of subjects with T2DM were lower than for males, and the associations between MA and T2DM, as well as all-cause mortality, were substantially weaker than for males.

We have identified numerous avenues for future studies: Replication of the MA prediction and evaluation of its clinical relevance in other cohorts with different population characteristics, prospective testing of the method on new subjects, and estimation of its sensitivity to realistic domain shift. The image-to-biomarker extraction method can easily be adapted to enable studies using subsets of body regions, tissues, organs, and image feature types, enabling a wide range of future medical studies. We aim to evaluate the proposed method on other interpretable precision medicine applications such as prediction of future diabetes and cardiovascular disease and oncological applications such as prediction of treatment response, treatment side-effects, overall survival from CT or PET/CT image data. The supervoxels could be constructed in a more data-driven manner, optimized to maximize performance on the prediction task and the image registration methodology might be further improved. Future studies could explore additional tissue groups, and tissue grouping being performed in the subject space rather than in reference space, reducing partial volume effects on the tissue group assignment.

Conclusion

We developed a novel image-to-biomarker conversion methodology and applied it to MA prediction from whole-body MR images in the UKBB. Linear and interpretable predictions were performed using localized and tissue-specific features resulting in state-of-the-art age prediction performance from whole-body MR images, outperforming a multi-organ and tissue segmentation based approach and previously reported CNN-based prediction results. The MAs were significantly associated with T2DM and all-cause mortality.

Acknowledgment

This research was enabled by grants from the Swedish Heart-Lung Foundation (20240402) and the Swedish Research Council (2016-01040, 2019-04756, 2023-03607), EXODIAB, and Uppsala Diabetes Center (UDC).

This research has been conducted using the UK Biobank Resource under Application Number 14237 and the Swedish National Infrastructure for Computing (SNIC, sens2019016). The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala University, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Authors J.Ö and J.K declare a patent-pending invention relevant for the proposed method. Authors A.B, Y.U, R. K. G, E.L, H.A declare no conflicts of interest.

References:

1) The Hallmarks of Aging, Carlos López-Otín , Maria A Blasco, Linda Partridge , Manuel Serrano, Guido Kroemer, Cell, 153(6), 1194-1217 (2013).
2) Biological age is superior to chronological age in predicting hospital mortality of the critically ill, Kwok M Ho, David J Morgan, Mason Johnstone, Cyrus Edibam, Internal and emergency medicine, 18(7), 2019-2028, (2023).
3) UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age, Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, Rory Collins, PLoS medicine 12.3 (2015): e1001779.
4) Identifying Morphological Indicators of Aging With Neural Networks on Large-Scale Whole-Body MRI, Taro Langner , Johan Wikström, Tomas Bjerner, Håkan Ahlström, and Joel Kullberg, IEEE TMI, (2019).
5) Atlas-Based Interpretable Age Prediction In Whole-Body MR Images, Sophie Starck, Yadunandan Vivekanand Kini, Jessica Johanna Maria Ritter, Rickmer Braren, Daniel Rueckert, Tamara Mueller, Journal of Machine Learning for Biomedical Imaging, 2 (2024).
6) Patch-based Brain Age Estimation from MR Images, Kyriaki-Margarita Bintsi, V. Baltatzis, A. Kolbeinsson, A. Hammers, D. Rueckert, MICCAI (2020).
7) Deep regression for biological age estimation in multiple organs: Investigations on 40,000 subjects of the UK Biobank, Veronika Ecker, Marcel Früh, Bin Yang, Sergios Gatidis, and Thomas Küstner. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2255-2259. IEEE, 2024.
8) Using deep learning to predict abdominal age from liver and pancreas magnetic resonance images, Le Goallec, Alan, Samuel Diai, Sasha Collin, Jean-Baptiste Prost, Théo Vincent, and Chirag J. Patel, Nature Communications 13, no. 1 (2022): 1979.
9) Validation of biomarkers of aging, Moqri, M., Herzog, C., Poganik, J. R., Ying, K., Justice, J. N., Belsky, D. W., ... & Ferrucci, L. (2024). Nature medicine, 30(2), 360-372.
10) Fast graph-cut based optimization for practical dense deformable registration of volume images. Simon Ekström, Filip Malmberg, Håkan Ahlström, Joel Kullberg, & Robin Strand, Computerized Medical Imaging and Graphics, 84, 101745 (2020).
11) Faster dense deformable image registration by utilizing both CPU and GPU. Simon Ekström, Martino Pilia, Joel Kullberg, Håkan Ahlström, Robin Strand, and Filip Malmberg. Journal of Medical Imaging 8, no. 1 (2021).
12) Deep learning-based age estimation from clinical Computed Tomography image data of the thorax and abdomen in the adult population. Bjarne Kerber, Tobias Hepp, Thomas Kuestner, and Sergios Gatidis. Plos one 18, no. 11 (2023): e0292993.
13) SLIC Superpixels Compared to State-of-the-Art Superpixel Methods, Radhakrishna Achanta *et al*
14) Imaging biomarkers of ageing: a review of artificial intelligence-based approaches for age estimation. Fridolin Haugg, Grace Lee, John He, Justin Johnson, Anna Zapaishchykova, Danielle S Bitterman,

Benjamin H Kann, Hugo J W L Aerts, Raymond H Mak, Review Lancet Healthy Longev. 2025 Jul;6(7):100728. doi: 10.1016/j.lanhl.2025.100728. Epub 2025 Jul 18.
15) deform, Simon Ekström, Github repository, github.com/simeks/deform, (commit 8c0651f).
16) Morphological age prediction interpretability maps from the UK Biobank MRI, Zenodo, 2025: https://doi.org/10.5281/zenodo.17313960.
17) VIBEsegmentator: Full Body MRI Segmentation for the NAKO and UK Biobank. Robert Graf, Paul-Sören Platzek, …, Daniel Rueckert, Jan Stefan Kirschke, arXiv preprint arXiv:2406.00125 (2024).
18) The design of SimpleITK. Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. Frontiers in neuroinformatics 7 (2013): 45.
19) UK Biobank Age Prediction (TS-TSSP), github.com/Radiological-Image-Analysis-Group-UU/ukbb_age_prediction.
20) Towards prediction of morphological heart age from computed tomography angiography, Johan Öfverstedt Elin Lundström Håkan Ahlström, Joel Kullberg, arXiv preprint, 2025.
21) Geroprotectors: a unified concept and screening approaches, Moskalev, Alexey, Elizaveta Chernyagina, Anna Kudryavtseva, and Mikhail Shaposhnikov, *Aging and disease* 8, no. 3 (2017): 354.
22) Shortcut learning in deep neural networks, Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann, Nature Machine Intelligence, 2(11), 665-673.
23) Convolutional neural networks: an overview and application in radiology, Yamashita, Rikiya, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi, Insights into imaging 9 (2018): 611-629.
24) Accurate brain age prediction with lightweight deep neural networks, Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, Stephen M. Smith, Medical image analysis, 68, (2021).
25) Toward explainable artificial intelligence for regression models: A methodological perspective, Letzgus, Simon, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Gregoire Montavon. IEEE Signal Processing Magazine 39, no. 4 (2022): 40-58.
26) Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. Ye E. Tian, Vanessa Cropley, Andrea B. Maier, Nicola T. Lautenschlager, Michael Breakspear, and Andrew Zalesky. Nature medicine 29, no. 5 (2023): 1221-1231.
27) Benchmarking saliency methods for chest X-ray interpretation, Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., ... & Rajpurkar, P, Nature Machine Intelligence, 4(10), 867-878 (2022).
28) Deep learning-based age estimation from chest CT scans, Azarfar, G., Ko, S. B., Adams, S. J., & Babyn, P. S., International Journal of Computer Assisted Radiology and Surgery, 19(1), 119-127 (2024).
29) AI-driven preclinical disease risk assessment using imaging in UK biobank, Seletkov, D., Starck, S., Mueller, T. T., Zhang, Y., Steinhelfer, L., Rueckert, D., & Braren, R. (2025). npj Digital Medicine, 8(1), 480.
30) FaceAge, a deep learning system to estimate biological age from face photographs to improve prognostication: a model development and validation study. Bontempi, D., Zalay, O., Bitterman, D. S., Birkbak, N., Shyr, D., Haugg, F., ... & Aerts, H. J. The Lancet Digital Health, (2025)

Supplementary material

A: Image registration

Image registration was performed using a deformable image registration method based on solving sequences of local-block (block-size 12 along all three dimensions) graph-cut optimization problems in a multi-resolution 6-level Gaussian pyramid framework (10, 11). The input consisted of normalized water fraction and fat fraction images, as well as AI-derived masks, obtained with VIBESegmentator, automatically delineating subcutaneous adipose tissue (SAT) and muscle tissue. The channel-wise sum of square differences was used as the objective function for the image registration (both for the intensity images and tissue masks). The water and fat fraction images were given a weight of 1, and the tissue masks were given a weight of 0.6. A spatial regularization term of 0.25 was selected based on preliminary experiments in order to enforce plausibility of the deformations, while still allowing enough flexibility to attain high anatomical overlap. 100 iterations were performed per level in the multi-resolution pyramid (or stopping early in case of no improvement).

The *deform (15)* package produced a displacement vector field representing the local voxel-wise correspondences between the reference subject and each of the other subjects. The contents of the individual subject FF volumes were then deformed into the reference space by applying inverse mapping with the displacement vector field to the corresponding FF volume with trilinear interpolation, bringing all FF volumes into the same sex-stratified template space. Additionally, the *deform* package provided the local Jacobian determinant (JD) for each voxel, which is a relative measure of voxel size, where a JD>1 indicates an expansion and a JD<1 indicates a contraction, in comparison to the template subject in that specific voxel, and where JD<0 indicates a fold of space.

B: Supervoxels

For the SLIC algorithm, the implementation in the scikit-image package (14) was used. The SLIC algorithm implementation is parametrized by the parameter n_{seg} (fewer than the original number of voxels), controlling the number of supervoxels that the algorithm aims to partition the image into, as illustrated in Fig. 2. A higher n_{seg} input results in a greater preservation of data with a higher spatial detail and an increased number of features, which in turn increases the noise of the features, the computational time and the size of the feature sets. Moreover, supervoxels are allowed to split or merge during the course of the algorithm's operation, so the exact number of supervoxels may differ from the provided n_{seg}.

C: Tissue-specific feature extraction

Formally, the features are defined as follows. Let x denote a voxel in the reference image space, S denote a supervoxel defined in the reference image space, and t denote the tissue group (e.g., lean tissue). $I(x)$ denotes the FF of voxel x, $JD(x)$ denotes the Jacobian Determinant of voxel x, $T_t(x)$ is a real-valued tissue weight between 0 and 1 for tissue group t and voxel x, $N_t(S)$ is the total weight of the voxels (also referred to as Voxel Count when used as a feature) in supervoxel S for tissue group t. The tissue-specific features f_t for tissue group t are then given by weighted averages, sums, and standard deviations, given by

$$f_t^{Mean\,FF}(S) = \frac{1}{N_t(S)} \sum_{x \in S} T_t(x) I(x) , \tag{1}$$

$$f_t^{Total\,VOL}(S) = \sum_{x \in S} T_t(x) JD(x) , \tag{2}$$

$$f_t^{Mean\,JacDet}(S) = \frac{1}{N_t(S)} \sum_{x \in S} T_t(x) JD(x) , \tag{3}$$

$$f_t^{SD\,FF}(S) = \sqrt{\frac{1}{N_t(S) - \frac{N_t^2(S)}{N_t(S)}} \sum_{x \in S} \left[T_t(x) \left(I(x) - f_t^{Mean\,FF}(S) \right) \right]^2} , \tag{4}$$

$$f_t^{SD\,JacDet}(S) = \sqrt{\frac{1}{N_t(S) - \frac{N_t^2(S)}{N_t(S)}} \sum_{x \in S} \left[T_t(x) JD(x) - \sum_{x \in S} f_t^{Mean\,JacDet}(S) \right]^2} , \tag{5}$$

where

$$N_t(S) = \sum_{x \in S} T_t(x), \tag{6}$$

$$N_t^2(S) = \sum_{x \in S} T_t(x)^2.$$

In this study, all tissue weights are binary, i.e., every voxel is assigned to a single tissue group. In general, the features are defined for non-binary tissue weights, thereby support the probabilistic assignments of voxels into tissue groups (e.g., a probabilistic model predicts that voxel x contains lean tissue with probability 0.6 and adipose tissue with probability 0.4).

The deformed FF maps were obtained by application of inverse mapping with tri-linear interpolation. JD fields were computed from the image registration deformation fields.

D: Feature pre-processing: z-score normalization, imputation of missing features, and dimensionality reduction

Given a feature f_i, its mean $\overline{f}$, standard deviation SD(f), and a limit c denoting how many standard deviations to clip the features at, the clipped z-scores were computed by

$$z_{f_i} = min\left[max\left(\frac{f_i - \overline{f}}{SD(f)}\right), -c), c\right].$$

In all experiments, we used c=3 based on preliminary experiments on the CV sets.

When using multiple tissue groups, some supervoxels may not contain enough voxels belonging to a group for the features to be well-defined (e.g., several supervoxels in the liver lacked adipose tissue voxels in all subjects). To account for this, features where 90% or more across the entire training set were undefined were dropped. Undefined features that were not dropped were imputed with a z-score of 0.

E: Details on the PCA analysis

The number of input feature dimensions was reduced to

$$n_{PCA} = min(0.9n_{training\ samples}, 0.9n_{features}, 3000) \quad (5)$$

principal components, limiting the number of components to fewer than the theoretical limit, after observing instabilities in the performance without this restriction. When a low number of PCA components is desired in order to limit the flexibility of the regression, we computed the univariate absolute Pearson correlation coefficient between each PCA component and the target variable (training set) and selected the $n_{Top-k\ PCA}$ components that exhibit the strongest association with the target variable (here CA).

The statistics used for z-score normalization and the PCA components were computed during model training from the training set features. The constant c was set to 3 (clipping the features at -3 and +3 standard deviations), a decision made from initial tests during cross-validation to obtain a good trade-off between preserving informative feature variation and removing noise and outliers.

F: Linear modelling

Linear regression models were selected during initial rounds of cross-validation experiments for having several beneficial properties compared to non-linear models, e.g., multi-layer perceptron models: 1) efficiency of model fitting and prediction on unseen data, 2) the same observed prediction performance or better than the deeper models tested, and 3) superior interpretability of the linear models over more complex models.

G: Debiasing of the MAs for the computation of age gaps

For 10 equidistant grid-points in the range between the minimum and the maximum age, we computed the moving average MA, with weights drawn from an unnormalized Laplace kernel with a bandwidth of 0.25 years, given by

$$w[C,\, r](x) \;=\; e^{-\frac{1}{0.25}|C(x)-r|},$$

where r is the grid-point (in years), C(x) denotes the CA of subject x. The updated MAs were then given by linear interpolation between the grid points. The correlation before and after bias correction is presented in Table G1.

Table G1: Effect of debiasing the 10 and 3000 component MAs. Before debiasing (pre), the correlation was strongly negative, and after debiasing (post), the correlation was near zero.

Components	Sex	r (Pre)	P-value (Pre)	r (Post)	P-value (Post)
10	Female	-0.607	0.0	-0.0067	0.352
	Male	-0.637	0.0	0.0096	0.195
3000	Female	-0.356	0.0	0.0343	$1.31*10^{-5}$
	Male	-0.358	0.0	0.0167	$3.33*10^{-6}$

H: Statistics

Bonferroni-adjusted P-values for the comparative tests between the 3-tissue configuration and the 1-tissue configuration, as well as between the 3-tissue configuration with supervoxels and the 3-tissue configuration with VIBESegmentator-derived masks, are presented in Table H1. Bonferroni-adjusted P-values for the comparative tests between the 3-tissue configuration and the 1-tissue configuration of VIBESegmentator-derived masks are presented in Table H2.

Table H1. Bonferroni-adjusted P-values for Wilcoxon signed-rank tests on absolute differences of predictions and reference CA between the 9 different models trained on the 9 different folds using cross-validation. The supervoxel experiments were performed with the 5000 supervoxel setup and 3000 PCA components. All adjusted p-values indicate statistically significant differences (threshold 0.05) except for fold 6 for males in the comparison between the 3-tissue configuration and the 1-tissue configuration.

Method	Sex	1	2	3	4	5	6	7	8	9
3-T SV and 1-T SV	Female	**$8.92*10^{-5}$**	**$9.87*10^{-6}$**	**$1.15*10^{-5}$**	**$1.74*10^{-4}$**	**$2.25*10^{-5}$**	**$4.63*10^{-6}$**	**$4.83*10^{-5}$**	**$3.82*10^{-3}$**	**$2.84*10^{-4}$**
	Male	**$7.05*10^{-5}$**	**$9.60*10^{-5}$**	**$1.66*10^{-3}$**	**$1.89*10^{-4}$**	**$2.72*10^{-4}$**	$9.02*10^{-1}$	**$8.70*10^{-3}$**	**$2.75*10^{-2}$**	**$1.15*10^{-3}$**
3-T SV and 3-T VIBESeg.	Female	**$4.10*10^{-88}$**	**$1.85*10^{-87}$**	**$1.20*10^{-89}$**	**$6.98*10^{-86}$**	**$2.15*10^{-88}$**	**$1.40*10^{-87}$**	**$6.38*10^{-85}$**	**$5.88*10^{-84}$**	**$2.88*10^{-85}$**
	Male	**$2.49*10^{-63}$**	**$3.18*10^{-64}$**	**$9.66*10^{-60}$**	**$2.71*10^{-66}$**	**$3.78*10^{-62}$**	**$1.09*10^{-59}$**	**$7.12*10^{-62}$**	**$6.49*10^{-64}$**	**$4.64*10^{-62}$**

Table H2. Bonferroni-adjusted P-values for Wilcoxon signed-rank tests on absolute differences of predictions and reference CA between the 9 different models trained on the 9 different folds using cross-validation. All adjusted p-values indicate statistically significant differences (threshold 0.05).

Method	Sex	1	2	3	4	5	6	7	8	9
3-T VIBESeg. and 1-T VIBESeg.	Female	**$2.83*10^{-5}$**	**$4.26*10^{-6}$**	**$4.00*10^{-5}$**	**$2.84*10^{-6}$**	**$4.81*10^{-6}$**	**$4.49*10^{-5}$**	**$6.68*10^{-6}$**	**$7.64*10^{-5}$**	**$8.71*10^{-6}$**
	Male	**$7.06*10^{-6}$**	**$4.83*10^{-7}$**	**$2.13*10^{-6}$**	**$1.68*10^{-5}$**	**$7.50*10^{-7}$**	**$1.01*10^{-6}$**	**$3.13*10^{-6}$**	**$1.60*10^{-5}$**	**$2.26*10^{-6}$**

I: Ablation studies

In Fig. I1, the cross-validation performance on age prediction as a function of the number of training subjects is presented.

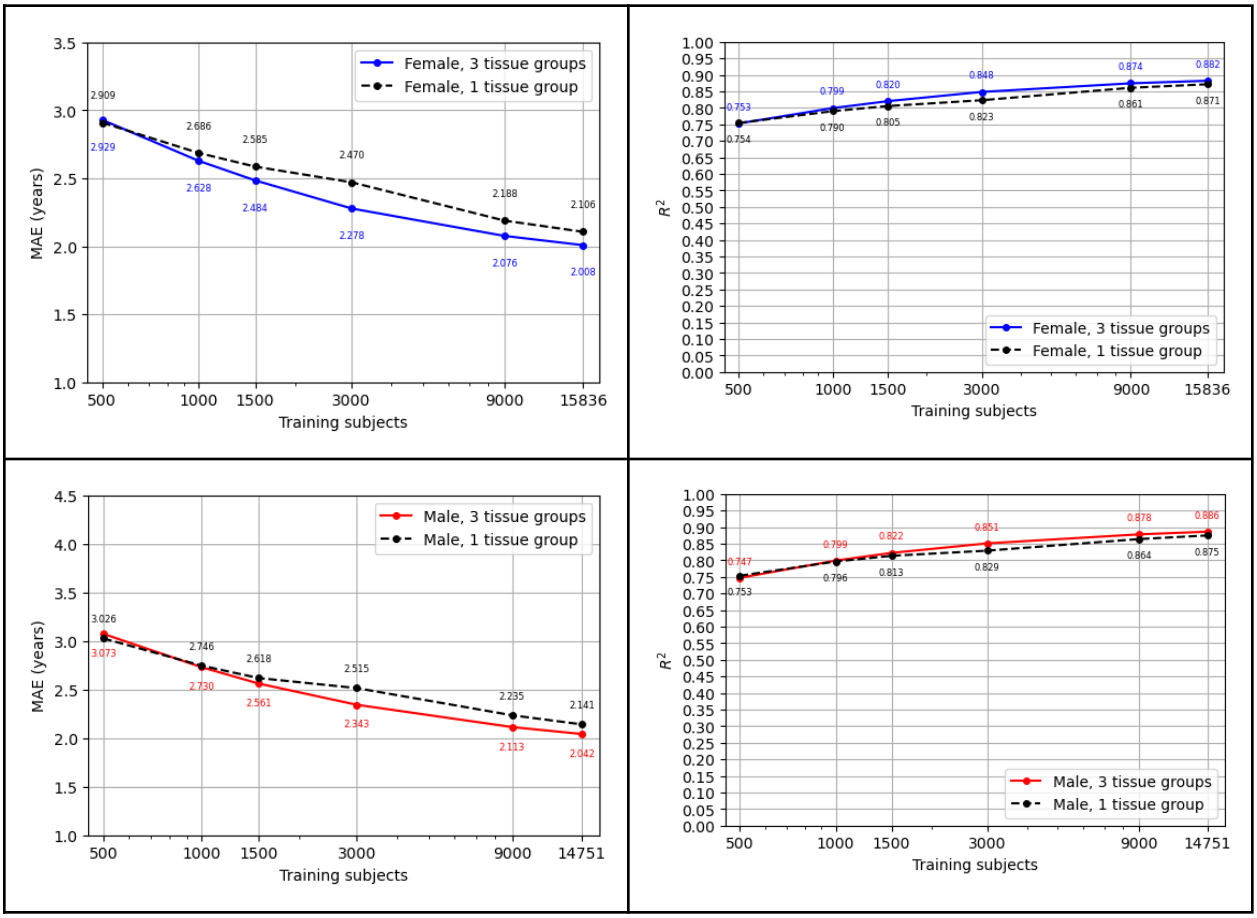

Figure I1: The effect of different numbers of training subjects on R^2 and MAE performance on age prediction across the cross-validation, comparing the 1-tissue group configuration with the 3-tissue group configuration. A higher R^2 and a lower MAE indicate better performance. The y-axis is scaled linearly, while the x-axis is scaled logarithmically. 5000 supervoxels were used for these experiments.

J: Chronological age prediction performance with 10 PCA components vs 3000 PCA components

Table J1 presents a comparison of the performance of CA prediction with 10 PCA components and 3000 PCA components.

Table J1: Performance of prediction of CA, evaluated with 10 and 3000 PCA components. The performance was measured using R^2 and MAE. Each cell presents the mean results of nine cross-validation models on the test set, with the standard deviation across models in parentheses.

PCA components	*Partition*	*R^2 Female*	*R^2 Male*	*MAE (years) Female*	*MAE (years) Male*
10	CV	0.635 (0.014)	0.596 (0.012)	3.579 (0.056)	3.894 (0.066)
3000	CV	**0.882 (0.004)**	**0.886 (0.003)**	**2.006 (0.028)**	**2.043 (0.022)**
10	Test set	0.620 (0.006)	0.611 (0.002)	3.591 (0.028)	3.919 (0.015)
3000	Test set	**0.884 (0.001)**	**0.892 (0.001)**	**1.951 (0.009)**	**2.057 (0.014)**

K: Association of the predicted MAs with diabetes and all-cause mortality

The results of the association of the age gap and prevalence of T2DM are presented in Fig. K1.

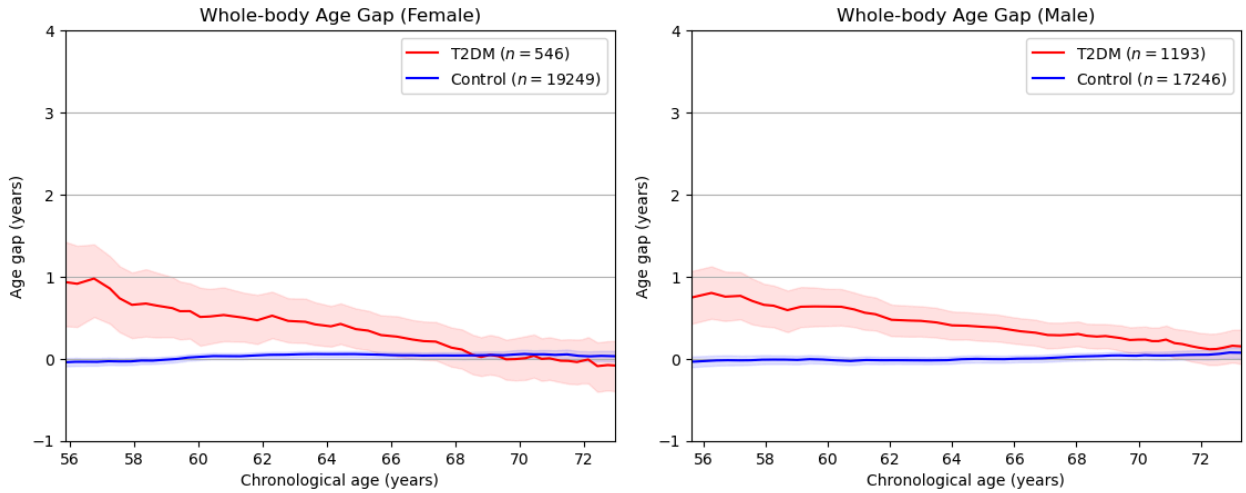

Fig K1: Association of type-2 diabetes (T2DM) and the predicted (and bias-corrected) MAs from 3000 PCA components (compared to 10 for the results in the main text), as a function of CA, computed as a sliding window with average bandwidth ± 10 years. The age gap was corrected to a value close to zero for all ages in the control group. The T2DM group exhibits a substantially higher mean age gap for all CAs. CIs were generated by bootstrapping with 1000 samples.

The results of the survival analysis for 3000 PCA components are presented as Kaplan-Meier plots in Fig K2.

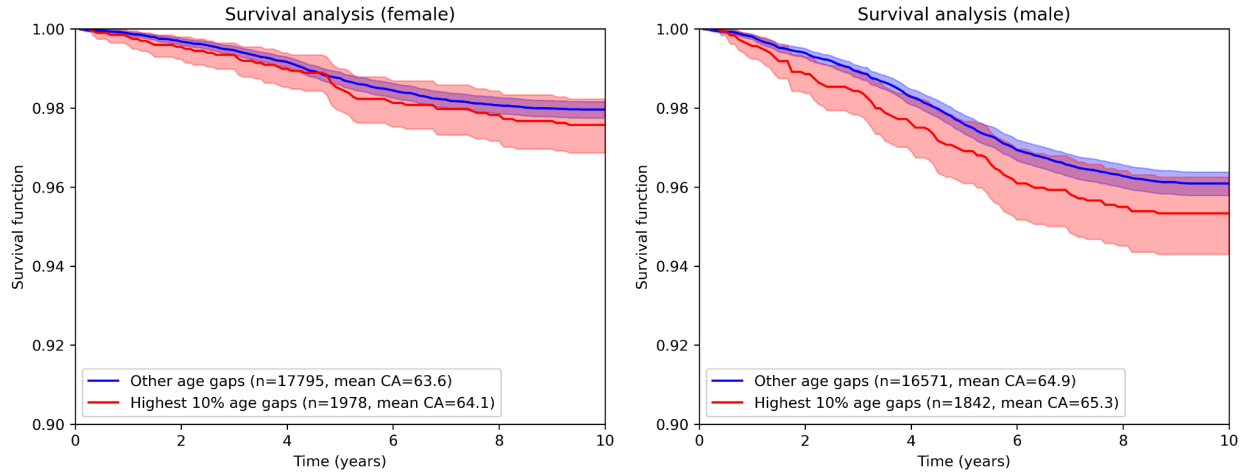

Fig K2: Survival analysis based on the predicted MAs for the model with 3000 PCA components. The highest 10% age gaps denote the group of subjects with an age gap larger than the top 10th percentile in each sub-cohort. The survival function models the number of subjects surviving beyond a given time point. 10-year or longer follow-up time is available for all analyzed subjects. CIs were generated by bootstrapping with 1000 samples.

L: Software, versions, and code sharing

SimpleITK, v2.3.1 (18), NumPy, v1.26.4 (L1), scikit-learn, v1.2.2, scikit-image, v0.22.0 (L2), and *deform,* v0.5.2 (10, 11, 15) were the main software packages and versions used for the image data analysis.

References

L1) Array programming with NumPy, CR Harris, et al
L2) scikit-image: Image processing in Python, Stefan van der Walt, et al

M: Tissue-wise and feature-wise age prediction performance

Female	Quality measure	Mean FF	Total Volume	Mean JacDet	Voxel Count	SD FF	SD JacDet
Lean tissue	MAE (y)	2.930 (0.012)	2.674 (0.0014)	2.758 (0.016)	2.870 (0.021)	3.123 (0.010)	3.143 (0.024)
	R2	0.749 (0.002)	0.788 (0.002)	0.773 (0.002)	0.754 (0.003)	0.711 (0.002)	0.705 (0.004)
Medium FF tissue	MAE (y)	2.557 (0.014)	2.495 (0.016)	2.614 (0.017)	2.474 (0.013)	2.651 (0.008)	2.878 (0.023)
	R2	0.805 (0.001)	0.815 (0.002)	0.793 (0.002)	0.816 (0.002)	0.791 (0.001)	0.752 (0.004)
Adipose tissue	MAE (y)	3.038 (0.022)	2.926 (0.020)	2.945 (0.015)	2.805 (0.015)	3.180 (0.012)	3.192 (0.014)
	R2	0.726 (0.004)	0.741 (0.002)	0.741 (0.003)	0.765 (0.003)	0.698 (0.002)	0.702 (0.003)

Male	Quality measure	Mean FF	Total Volume	Mean JacDet	Voxel Count	SD FF	SD JacDet
Lean tissue	MAE (y)	2.970 (0.016)	2.673 (0.012)	2.794 (0.012)	2.942 (0.013)	3.060 (0.011)	3.188 (0.023)
	R2	0.773 (0.001)	0.816 (0.001)	0.794 (0.002)	0.778 (0.002)	0.760 (0.002)	0.735 (0.003)
Medium FF tissue	MAE (y)	2.690 (0.018)	2.655 (0.016)	2.724 (0.010)	2.553 (0.019)	2.759 (0.023)	3.011 (0.021)
	R2	0.813 (0.002)	0.816 (0.002)	0.810 (0.001)	0.831 (0.002)	0.799 (0.003)	0.765 (0.003)
Adipose tissue	MAE (y)	3.199 (0.019)	3.224 (0.013)	3.162 (0.019)	3.022 (0.018)	3.453 (0.024)	3.488 (0.015)
	R2	0.734 (0.003)	0.732 (0.002)	0.745 (0.003)	0.758 (0.002)	0.690 (0.004)	0.687 (0.003)

N: Computational pipeline and software runtime

The main steps performed by the analysis pipeline from raw images are sequentially in order:
1) Stitching of the images from the different MR imaging stations.
2) Computation of the FF, WF and sum images.
3) Segmentation using VIBESegmentator using the sum images as input.
4) Registration of the image to the reference image.
5) Computation of the deformed FF image and the JacDet image.
6) Tissue group assignment based on deformed FF images.
7) Supervoxel-based feature extraction by supervoxel, tissue group, feature combination.
8) Z-score standardization and filtering of features.
9) PCA dimensionality reduction.
10) Application of the linear model to the PCA-components.

The approximate runtime of the full pipeline on a Intel i9-13900K, 64 GB RAM machine for a new subject is 5 minutes. The majority of the runtime is due to the inter-subject image registration, and to a lesser part the stitching, segmentation and feature extraction. The PCA-based dimensionality reduction and application of the linear model is almost a negligible part of the runtime.

O: Interpretability studies and interpretability maps

The interpretability studies involve the creation of visual interpretability maps, showing the relationships between the features and the target variable used by the models overlayed on the standardized supervoxels in the reference image space where they were computed. Starting with the coefficients of the linear models taking the PCA-components as inputs, we map back their relationship to the original input features, by multiplication of the PCA loadings with the model coefficients, as given by

$$a_i = \sum_j w_{i,j} \beta_j$$

where $w_{i,j}$ denotes the PCA loadings for feature i and component j, β_j denotes the linear coefficient from PCA component j and the target variable.

For visualization, these feature coefficients are normalized to the -1 to +1 range by division by the maximum absolute value out of all a_i.

To simplify the viewing of the interpretability maps, each association g_i is plotted in the supervoxel it was computed from, and only in the voxels belonging to the tissue group

corresponding to feature i in the reference subject. This enables the simultaneous viewing of the associations from all tissue groups in a single map.