MIMIR: Deep Regression for Automated Analysis of UK Biobank Body MRI

Taro Langner, Andrés Martínez Mora, Robin Strand, Håkan Ahlström, and Joel Kullberg

***Abstract*—UK Biobank (UKB) is conducting a large-scale study of more than half a million volunteers, collecting health-related information on genetics, lifestyle, blood biochemistry, and more. Medical imaging furthermore targets 100,000 subjects, with 70,000 follow-up sessions, enabling measurements of organs, muscle, and body composition. With up to 170,000 mounting MR images, various methodologies are accordingly engaged in large-scale image analysis. This work presents an experimental inference engine that can automatically predict a comprehensive profile of subject metadata from UKB neck-to-knee body MRI. In cross-validation, it accurately inferred baseline characteristics such as age, height, weight, and sex, but also emulated measurements of body composition by DXA, organ volumes, and abstract properties like grip strength, pulse rate, and type 2 diabetic status (AUC: 0.866). The proposed system can automatically analyze thousands of subjects within hours and provide individual confidence intervals. The underlying methodology is based on convolutional neural networks for image-based mean-variance regression on two-dimensional representations of the MRI data. This work aims to make the proposed system available for free to researchers, who can use it to obtain fast and fully-automated estimates of 72 different measurements immediately upon release of new UK Biobank image data.**

I. Introduction

UK Biobank (UKB) has shared an extensive record of health-related data for more than half a million volunteers with over 2,000 research projects. Starting in 2014, medical imaging was initiated for 100,000 participants with several modalities such as DXA, ultrasound, and MRI [1]. Beyond collecting data on genetics, lifestyle, and biochemistry of blood and urine, ongoing research is accordingly engaged in large-scale analysis of the extensive acquired imaging data.

Information on body composition [2] and liver fat content [3] can be extracted from these images, with important implications for metabolic and cardiovascular disease [4]. Although the required image analysis can involve human labor, fully-automated approaches more recently introduced neural networks for segmentation of the liver [5], pancreas [6], [7], kidneys [8], and various other muscles, organs, and tissues within this study [9], [10]. Many of these techniques target the UKB neck-to-knee body MRI, which can represent almost all of human anatomy in one comprehensive image. Lean and adipose tissue can be clearly distinguished in these images, based on a two-point Dixon technique that acquires separate, volumetric water and fat signal images.

Department of Surgical Sciences, Uppsala University, Sweden (T.L., AM.M. R.S., H.A., J.K.); Department of Information Technology, Uppsala University, Sweden (R.S.); and Antaros Medical AB, Sweden (H.A., J.K.).

Address correspondence to T.L. (e-mail: *taro.langner@surgsci.uu.se*).

Despite these developments, many relevant measurements only cover a small fraction of the 40,000 images that have been available for more than a year. The evaluation of the image data, quality control, and sharing of measurements is accordingly lagging far behind the image acquisition. Researchers with access to the data, who are conducting correlation analyses to genetics, lifestyle, and blood biochemistry, but also longitudinal developments related to aging, are accordingly confined to comparably small sample sizes. Months or years can pass until measurements for newly released images become available, and this backlog can only be expected to grow over time as more images are released [1].

This work presents a freely available software tool that can infer a wide range of these measurements automatically within hours.[1] It can process UKB neck-to-knee body MRI, or image data from other studies that reproduce the protocol on similar demographics, and predict 72 measurements together with individual confidence intervals. In this work, the prediction model was created and retrospectively validated against existing data. The value of this system consists in making automated measurements conveniently available to researchers, months or years ahead of time, almost immediately whenever new UKB images are released.

II. Methods

Based on convolutional neural networks, ***M**edical **I**nference on **M**agnetic resonance images with **I**mage-based **R**egression* (MIMIR) performs an image-based, deep regression [11]. A similar methodology was independently proposed for UKB brain MRI to estimate human age, which subsequently enabled correlation studies to genetics data [12].

The neural networks process UKB neck-to-knee body MRI, compressed into a two-dimensional format as seen in Fig. 1, and predicts the mean and variance of a Gaussian probability distribution over each given target measurement [13]. As a result, it can provide both a point estimate μ for the measurement itself and a heteroscedastic variance σ^2, which can expresses aleatoric uncertainty and describe confidence- or prediction intervals [14]. This methodology was previously proposed for six measurements of body composition on the same data [15]. Here, neural network instances were not formed into ensembles, however, in favor of speed and convenience. The targets were furthermore expanded to 72 different measurements, ranging from age, height, weight, and sex over body composition and organ volumes to experimental properties like grip strength, pulse rate, and type 2 diabetic status [16].

[1] github.com/tarolangner/ukb_mimir

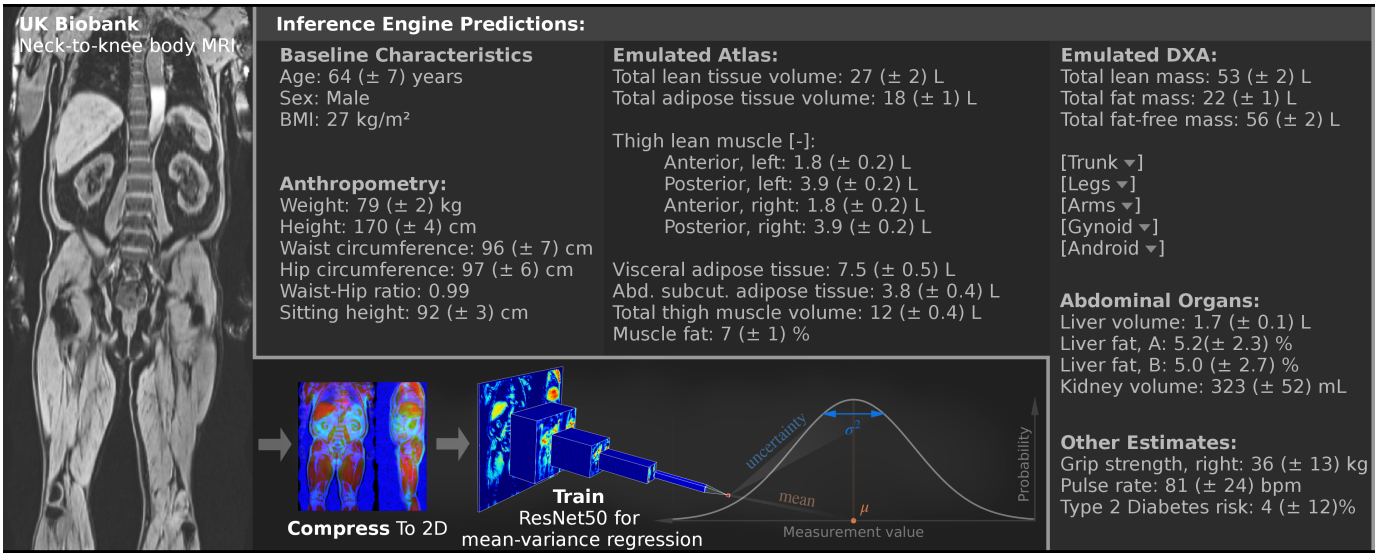

Fig. 1. MIMIR is an experimental inference engine for prediction of measurements and metadata from UKB neck-to-knee body MRI with image-based deep regression. Here, image data for one subject (left) was compressed and processed as part of the validation set. The listed measurements were inferred, together with 95% confidence intervals based on the mean and variance predicted by a ResNet50 convolutional neural network (bottom part).

Neck-to-knee body MRI was shared by UKB for about 40,000 men and women aged 44-82 (mean 63) years, BMI 14-62 (mean 27) kg/m^2 and 95% self-reported white British ethnicity (see also [1], [2]). Following visual inspection for exclusion of artifacts, severe pathologies, and other anomalies, 38,916 subjects remained for the experiments [15].

From the UKB metadata, 72 fields were selected as regression targets. They were grouped into four modules, for each of which one network instance with one or more outputs was trained: Body composition, abdominal organs, anthropometric/experimental estimates, and age. The available reference values for these regression targets originate from previously shared atlas-based segmentation results [2], manual analyses [3], DXA imaging [1], and prior neural network segmentations of the liver and kidneys [8]. Across all subjects, 73% of these values were not available through UKB. Estimating these missing measurements and providing them automatically for future UKB images was the main motivation for this work.

A stratified, 10-fold cross-validation split was generated by grouping all subjects into ten even subsets, each of which in turn served for validation of a neural network trained on the data of all remaining sets. This pretrained ResNet50 [17] with a specialized mean-variance loss function [14] is visualized in Fig. 1. All samples with at least one known ground truth value were used by setting the loss for missing values to zero. The network was trained in PyTorch with the Adam optimizer, batch size 32, and augmentation by random translations. After training for 8,000 iterations, the learning rate was reduced from $5e-5$ to $5e-6$ for another 2,000 iterations.

The predictions were evaluated with the intraclass correlation coefficient (ICC) using a two-way random, single measures, absolute agreement definition. For this metric, the reliability can be considered good for values above 0.75 and excellent for those above 0.90, with a maximum of 1.0 [18]. Additionally, the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE) and the area under curve of the receiver operating characteristic curve (AUC-ROC) are provided. The uncertainty estimates are known to often underestimate the true prediction errors [19], and were therefore calibrated post-hoc with scaling factors determined on the validation data [15].

III. Results and Discussion

The cross-validation results are summarized in Fig. 2. The evaluation yielded mean absolute errors (MAE) of 2.6 years for estimation of chronological age, 1.8 cm for height, 0.9 kg for bodyweight, and correct identification of sex in all but 14 of 38,874 subjects (of whom five differed in registered vs genetic sex). Relative errors of 7% for parenchymal kidney volume, less than 5% for several body composition measurements, and 3% for total liver volume were incurred. Liver fat content was predicted with $R^2 = 0.955$ and probable type 2 diabetics [16] were identified with an AUC-ROC of 0.866. The inference engine can process thousands of subjects within minutes on an Nvidia RTX 2080 Ti with 11GB, and extensive tables list detailed metrics for all 72 targets online[2].

Several limitations apply. No independent test set was examined, and so the trained networks should not be expected to correctly process arbitrary MRI data. Generalization is likely restricted to images that resemble the UKB training data in protocol, imaging device, and demographics. Out-of-domain tasks would likely require hundreds of subjects for renewed training [11]. However, robust performance within UKB can likely be expected, as prior work saw similar systems match or exceed the cross-validation performance on withheld UKB subjects [15]. The predicted uncertainties underestimate the prediction errors, especially since no ensembling was used. However, the scaling factors likely resolve this issue, and may indeed yield somewhat conservative confidence intervals.

[2]github.com/tarolangner/ukb_mimir

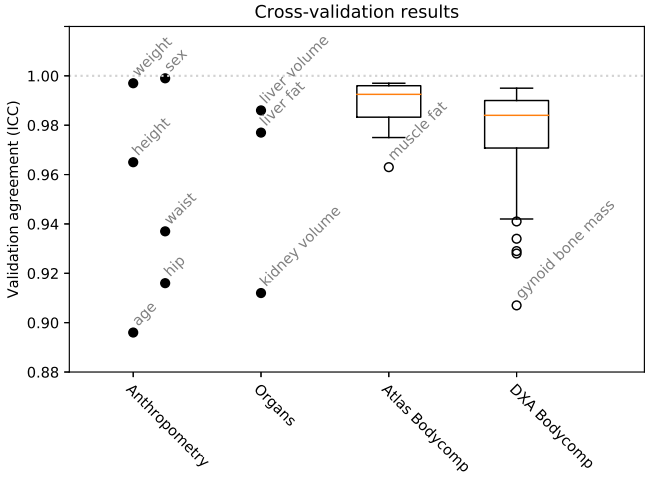

Fig. 2. Agreement between predictions and reference in cross-validation, expressed as intraclass correlation coefficient (ICC). Note that not all targets fall within the bounds of this plot.

Not all metadata can be inferred with the proposed approach, as the required information may be lost by preprocessing or absent from the MRI data. Other measurements are clinically trivial, such as age, sex, and weight. For some, alternative techniques may be superior, such as liver fat measurements from the dedicated UKB liver MRI with more accurate fat fraction values. Here, the projected, two-dimensional input format also limits the accuracy of organ measurements, with the proposed kidney volume measurements almost doubling the error previously achieved by segmentation of axial slices [8]. Finally, MRI-based diagnosis of type 2 diabetic status may not be clinically viable, with a specificity of 0.965 and sensitivity of only 0.250 here. Prior work classified it using convolutional neural networks with similar accuracy [20].

Future work may explore fully volumetric processing and other imaging protocols, which could leverage more, potentially critical information. However, several metrics can already be accurately inferred in the proposed way, such as liver volume and body composition metrics with estimated errors below 5%. With additional metadata becoming available, such as hormone levels, disease outcomes, and mortality statistics, more modules could be added in the future. The networks already trained in this work may provide benefit for independent studies that replicate the used imaging protocol, but also hold potential for transfer learning.

This work provides an implementation that can conveniently provide image-derived phenotypes at large scale for prototyping in correlation studies, months or years before the reference measurements will be available. Over the coming years, it will stand ready for fully-automated analysis of more than 100,000 upcoming UKB images that are yet to be released.

IV. Acknowledgments

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

References

[1] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, *et al.*, "The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[2] J. West, O. Dahlqvist Leinhard, T. Romu, R. Collins, S. Garratt, J. D. Bell, M. Borga, and L. Thomas, "Feasibility of MR-Based Body Composition Analysis in Large Scale Population Studies," *PLoS ONE*, vol. 11, Sept. 2016.

[3] H. R. Wilman, M. Kelly, S. Garratt, P. M. Matthews, M. Milanesi, A. Herlihy, M. Gyngell, S. Neubauer, J. D. Bell, R. Banerjee, *et al.*, "Characterisation of liver fat in the uk biobank cohort," *PloS one*, vol. 12, no. 2, p. e0172921, 2017.

[4] J. Linge, M. Borga, J. West, T. Tuthill, M. R. Miller, A. Dumitriu, E. L. Thomas, T. Romu, P. Tunón, J. D. Bell, *et al.*, "Body composition profiling in the uk biobank imaging study," *Obesity*, vol. 26, no. 11, pp. 1785–1795, 2018.

[5] B. Irving, C. Hutton, A. Dennis, S. Vikal, M. Mavar, M. Kelly, and J. M. Brady, "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.

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

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

[8] T. Langner, A. Östling, L. Maldonis, A. Karlsson, D. Olmo, D. Lindgren, A. Wallin, L. Lundin, R. Strand, H. Ahlström, *et al.*, "Kidney segmentation in neck-to-knee body mri of 40,000 uk biobank participants," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

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

[10] T. Kart, M. Fischer, T. Küstner, T. Hepp, F. Bamberg, S. Winzeck, B. Glocker, D. Rueckert, and S. Gatidis, "Deep learning–based automated abdominal organ segmentation in the uk biobank and german national cohort magnetic resonance imaging studies," *Investigative Radiology*, 2021.

[11] T. Langner, R. Strand, H. Ahlström, and J. Kullberg, "Deep regression for uncertainty-aware and interpretable analysis of large-scale body mri," 2021.

[12] B. A. Jónsson, G. Bjornsdottir, T. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson, "Brain age prediction using deep learning uncovers associated sequence variants," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, pp. 6402–6413, 2017.

[14] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in neural information processing systems*, pp. 5574–5584, 2017.

[15] T. Langner, F. K. Gustafsson, B. Avelin, R. Strand, H. Ahlström, and J. Kullberg, "Uncertainty-aware body composition analysis with deep regression ensembles on uk biobank mri," *arXiv preprint arXiv:2101.06963*, 2021.

[16] S. V. Eastwood, R. Mathur, M. Atkinson, S. Brophy, C. Sudlow, R. Flaig, S. de Lusignan, N. Allen, and N. Chaturvedi, "Algorithms for the capture and adjudication of prevalent and incident diabetes in uk biobank," *PloS one*, vol. 11, no. 9, p. e0162388, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[18] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv preprint arXiv:1706.04599*, 2017.

[20] R. Wagner, B. Dietz, J. Machann, P. Schwab, J. K. Dienes, S. Reichert, A. L. Birkenfeld, H.-U. Haering, F. Schick, N. Stefan, *et al.*, "102-or: Detection of diabetes from whole-body magnetic resonance imaging using deep learning," 2020.