

Population-Scale Segmentation of Penile Tissue in Dixon MRI using Deep Learning for Quantitative Phenotyping in Male Reproductive Health

Jan Ernsting^{1,2,3,4}[0000-0003-4649-2111]^{*}, Gunnar Paul Kordes⁵[0009-0003-7501-313X]^{**}, Nils Johannaber⁵[0009-0008-4676-5726], Lynn Ogoniak⁵[0000-0002-2739-0251], Wolfgang Roll⁵[0000-0002-3746-9393], Tim Hahn^{3,4}[0000-0001-6541-3795], Alexander Siegfried Busch⁵[0000-0003-4417-569X], and Benjamin Risse^{1,2}[0000-0001-5691-4029]^{* * *}

¹ Institute for Geoinformatics, University of Münster, Münster, Germany

² Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany

³ University of Münster, Institute for Translational Psychiatry, Münster, Germany

⁴ Institute for Machine Learning in Medicine (focus area psychiatry), University Münster, Münster, Germany

⁵ Department of General Pediatrics, University of Münster, Münster, Germany

Abstract. Penile measurement is clinically relevant across male reproductive and urogenital health, including conditions such as micropenis, congenital and endocrine disorders, and sexual or urinary dysfunction. However, quantitative assessment of penile size has relied mainly on external length or circumference measurements, which are difficult to standardize, sensitive to measurement conditions, and unable to capture the internal portion of the penis. MRI enables volumetric assessment of the whole penis in vivo, but automated segmentation has not previously been established at population scale. Automated whole-organ volumetry would enable high-throughput phenotyping for multi-omics and clinical studies of male reproductive disease.

Here, we present a deep learning framework for whole-penis segmentation in multi-channel Dixon MRI. Using a newly curated expert-annotated training dataset ($n = 145$ subjects; 13,050 annotated slices) and a double-annotated independent test benchmark ($n = 24$ subjects; 2,160 double-annotated slices), we optimized a 3D nnU-Net architecture. The model achieved a 5-fold cross-validation Dice score of 0.90 and performed at observer-level accuracy on the independent test set (Dice: 0.92; Hausdorff distance: 3.58).

We deployed the model in 34,412 UK Biobank participants, enabling automated quantification of total penile tissue, including both external

^{*} Corresponding author: j.ernsting@uni-muenster.de. Equal contribution.

^{**} Equal contribution.

^{***} b.risse@uni-muenster.de

and internal components. Longitudinal evaluation in 2,282 men demonstrated high inter-session reproducibility ($r = 0.86$). This framework establishes a reproducible and population-scalable method for MRI-based assessment of penile anatomy and provides an open technical resource for future studies in urological imaging and male reproductive health. The trained model weights will be publicly released.

Keywords: Penis Segmentation · Dixon MRI · UK Biobank · nnU-Net

1 Introduction

Male genital anatomy exhibits substantial inter-individual variation in shape and size, which may be influenced by developmental, hormonal, metabolic, and aging-related processes. Quantitative characterization is therefore relevant for biomedical research and clinical medicine, supporting the assessment of anatomical variation and abnormalities. However, standardized large-scale assessment of penile morphology remains limited. Existing population-level evidence relies predominantly on observer-dependent external measurements [11] that do not capture the proximal penile root, comprising the crura and bulb, despite its anatomical and clinical relevance to total erectile tissue [8]. This limitation is especially relevant for conditions affecting penile development or presentation, such as micropenis, buried penis, and androgen insensitivity syndrome, where whole-organ morphology may provide information beyond external dimensions alone [6].

Magnetic resonance imaging (MRI) offers the possibility of non-invasive three-dimensional characterization of penile anatomy, enabling visualization of both external and internal components. In particular, large population imaging resources such as the UK Biobank provide standardized Dixon MRI acquisitions that create unprecedented opportunities for systematic morphometric analysis across thousands of participants. However, accurate penis segmentation is complicated by the distinct challenges of the external and proximal penile compartments: the pendulous penis shows substantial variation in shape and position across scans, whereas the penile root is anatomically embedded within the perineum and closely bordered by adjacent muscles and pelvic soft tissues. These properties make whole-organ delineation more complex than segmentation of many spatially constrained internal organs.

Deep learning has become a key methodology for medical image segmentation and has enabled large-scale quantitative phenotyping in biomedical imaging. While recent frameworks have successfully segmented thoracic, abdominal, and pelvic structures [5], automated segmentation of penile tissue remains unexplored. To address this gap, we introduce a deep learning-based framework for population-scale penis segmentation in UK Biobank Dixon MRI.

The model is available here for download: <https://zivgitlab.uni-muenster.de/cvmls/population-scale-penile-tissue-segmentation-in-dixon-mri>.

1.1 Contribution

The primary contributions of this work are threefold: First, we introduce a novel, manually annotated dataset for penile MRI segmentation derived from the large-scale UK Biobank cohort. Second, we provide a highly reliable evaluation benchmark consisting of an independent test set independently annotated and verified by two domain experts, ensuring a robust ground truth for future comparative studies. Third, to facilitate reproducibility and accelerate future research in automated pelvic imaging, we publicly release our finalized deep learning model configurations and trained weights.

We deliberately build on the standard nnU-Net framework rather than proposing a new architecture. Our contribution is therefore one of application and clinical translation rather than algorithmic novelty.

2 Related Works

2.1 Segmentation in MRI data

Deep learning has fundamentally transformed semantic segmentation in medical imaging. Foremost among segmentation frameworks suited for medical data is the nnU-Net framework [7], which has established state-of-the-art benchmarks by automating data preprocessing, network topology configuration, and training hyperparameters based on dataset properties.

A notable advancement in this domain is VIBESegmentator [5], which was developed to address automated segmentation in volumetric Dixon MRI used in large-scale cohorts such as NAKO and UK Biobank for whole-body MRI segmentation. Tools like VIBESegmentator demonstrate the feasibility of processing high-throughput, multi-station abdominal and pelvic datasets to extract reliable morphological segmentations.

However, despite these advances in overall body composition and tissue segmentation, highly localized structures within the male pelvic region remain underrepresented. Addressing this gap, Ernsting et al. [3] recently demonstrated the utility of targeted deep learning models for male reproductive anatomy, specifically focusing on testis segmentation within the UK Biobank. While their work demonstrates usage of Dixon MRI data to segment small pelvic structures, the segmentation of penile anatomy remains an unaddressed challenge due to high structural variability, complex surrounding boundaries, and a lack of expert-annotated training data. This study directly extends previous efforts by introducing a dedicated framework and expert-verified benchmark for automated penis segmentation.

2.2 Evaluation of penile size and morphology

In clinical routine, assessment of penile size is performed for several distinct indications, including the evaluation of differences or disorders of sex development, congenital genital anomalies such as hypospadias or buried penis, endocrine disorders associated with impaired androgen production or action, the

diagnosis of micropenis, counselling of men with perceived penile shortness, evaluation of erectile and acquired structural abnormalities, and surgical planning and follow-up. Conventional methods primarily rely on anthropometric measurements, including flaccid penile length, stretched penile length, erect length, and circumference [11]. However, these clinical metrics remain substantially method-dependent, as measurements vary according to anatomical landmarks, stretching force, and observer experience, introducing considerable inter-observer variability and limiting comparability across studies [2]. Furthermore, conventional measurements are restricted to external dimensions and do not capture the complete anatomical and volumetric extent of the penis. Ultrasonography provides additional structural information but is mainly applied for selected anatomical or hemodynamic assessments and lacks a standardized framework for whole-organ quantification. MRI, although not routinely used for penile size evaluation, enables volumetric visualization of the complete penile anatomy, including both external and internal structures such as the glans, corpora cavernosa, corpus spongiosum with bulb, and crura [8]. Consequently, MRI-based AI segmentation represents a promising approach for standardized, observer-independent quantification of total penile tissue volume and morphology. While AI-assisted contouring of individual penile structures has shown feasibility in MRI-based studies [1,10], comprehensive penis segmentation remains largely unexplored.

3 Methods

3.1 Dataset

Data for this study were obtained from the UK Biobank, a large-scale, prospective biomedical database and research resource. Eligibility for the current analysis was restricted to male participants who completed the abdominal MRI protocol from 2014 onward. To ensure data integrity, the cohort was filtered to include only individuals with concordant self-reported and genetically inferred sex. This yielded a primary study population of 34,600 male participants aged 45–87 years (mean: 67, standard deviation [SD]: 8). Within this cohort, a subset of 2,315 participants underwent repeat imaging assessments from 2019 onward (1–10 years time interval; mean: 5 years; SD: 2 years), providing a longitudinal follow-up subcohort. Of the 34,600 eligible participants, 107 were excluded for incomplete acquisition of all four modalities and 81 for shape mismatch across modalities, yielding a final sample of 34,412 participants for the penis segmentation task. For model training, only the initial imaging sessions were considered.

The Dixon MRI data from the UK Biobank capture complementary anatomical structures across six spatial imaging slabs extending from the neck to the knee, and covering a total height of 1.10 meters [12]. This MRI technique generates four distinct volumetric reconstructions per slab: water, fat, in-phase, and out-of-phase images. For image processing and analysis, the original DICOM files were converted into the NIfTI format.

3.2 Preprocessing

Since the relevant anatomy spanned multiple imaging slabs, a preprocessing pipeline was implemented to combine the relevant acquisitions. Specifically, the two lowermost caudal imaging slabs of all four sequence modalities (water, fat, in-phase, and opposed-phase) were stitched together to form a continuous volumetric field of view. Stitching was performed by header-based spatial assembly, with a 2-voxel margin removed from both slab edges along the z-axis and voxel intensities averaged across overlapping regions to smooth intensity transitions between acquisition stations [9].

3.3 Annotation

Manual annotations were performed using 3D Slicer (version 5.10.0; <https://www.slicer.org>) [4] on opposed-phase sequences, ensuring that the penis was completely visible within the field of view across all included scans. The total dataset consisting of 169 randomly selected images from the UK Biobank dataset was partitioned into a training set ($n = 145$) and a test set ($n = 24$). The training set was divided between two independent readers, where Annotator 1 segmented 108 cases (9,720 annotated axial slices) and Annotator 2 segmented 37 cases (3,330 annotated axial slices). Conversely, the test set consisted of 24 MRIs (2,160 double-annotated axial slices) dedicated to double annotation, where both Annotator 1 and Annotator 2 independently segmented the same cases to facilitate rigorous inter-observer validation. These annotations were used to measure inter-annotator agreement and define the baseline for human annotation performance.

To resolve discrepancies and break ties between these double segmentations for final model evaluation, a third imaging expert blindly reviewed and selected the superior annotation for each case. This final set of 24 images is the test set used for model evaluation. We chose expert selection over voxel-wise fusion because averaging two masks that disagree mainly at ambiguous imaged tissue can yield boundaries rejected by both annotators. Selection preserves an internally consistent and expert-endorsed segmentation. We note this privileges anatomical plausibility as judged by a third expert over a fused consensus, and may retain annotator-specific conventions of the selected annotation.

3.4 Model

For the segmentation of penile structures, we employed the nnU-Net framework [7], adapting the methodological pipeline proposed by Ernsting et al. [3].

We trained and compared both the 2D and 3D full-resolution configurations of the architecture. The network architectures were configured to accept a four-channel input consisting of in-phase, opposed-phase, water, and fat images, which were mapped to the corresponding target binary segmentation masks.

Training was conducted strictly adhering to the default nnU-Net configuration.

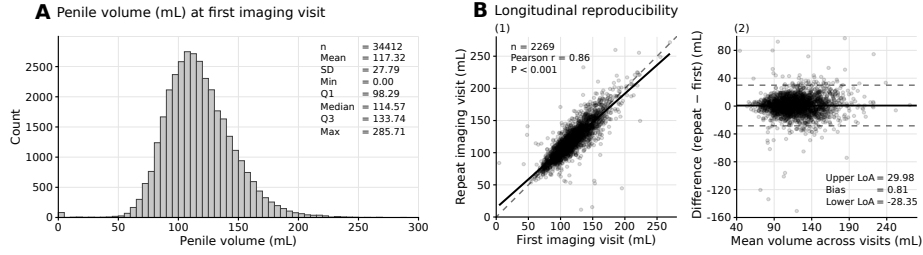


Fig. 1. (A) Predicted penile volume in milliliters (mL) for $n = 34,412$ UK Biobank subjects. Predictions with 0 mL were manually reviewed and identified as imaging errors, i.e. the penis not being visible in the merged slabs. (B) Longitudinal reproducibility of measurements across two imaging sessions (mean interval = 5 years; SD = 2 years) after excluding men with a 0 mL prediction in either session, leaving $n = 2,269$. (1) Scatterplot of first and repeat measurements with Pearson correlation and fitted linear regression. (2) Bland–Altman plot showing mean bias and 95% limits of agreement.

3.5 Training

Model training and inference were executed on a computing system equipped with an AMD EPYC 9124 CPU and an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM).

The experimental pipeline was structured into three phases: First, internal model training and architectural tuning were conducted using a 5-fold cross-validation scheme on the training dataset. Second, to evaluate generalization performance, model inference was executed on the test dataset. Third, the model was deployed to perform inference across the 34,412-participant cohort, generating the segmentation masks for subsequent large-scale population analysis.

4 Results

4.1 Human performance

To establish a baseline for human performance and evaluate inter-observer variability, the 24 holdout images in the test dataset were independently segmented by two domain experts. The spatial overlap between the two manual annotations yielded a mean Dice similarity coefficient (DSC) of 0.82 SD= 0.02 (median: 0.82). To evaluate boundary distances, the Hausdorff distance was calculated, revealing a mean boundary separation of 9.54 SD= 3.54 mm (median: 9.38 mm).

4.2 5-Fold Cross-Validation results

Both 2D and 3D network configurations were evaluated to determine the optimal dimensionality for the segmentation task.

The 3D nnU-Net configuration demonstrated superior performance and stability across all folds, achieving a mean DSC of 0.90 SD= 0.002 (median: 0.90).

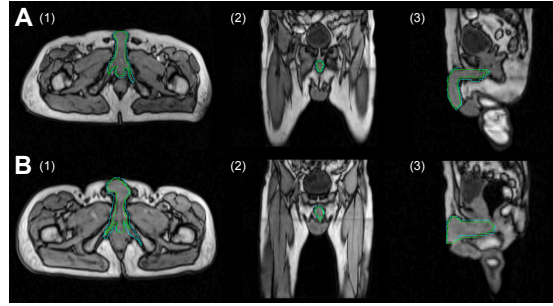


Fig. 2. Example annotations of penile tissue in UK Biobank Dixon MRI opposed-phase acquisitions using merged imaging slabs. Axial (1), coronal (2), and sagittal (3) views are shown. Green annotations indicate Annotator 1, and blue annotations indicate Annotator 2. Inter-annotator differences are primarily located in the proximal penile root, including the crura and bulb. Images reproduced by kind permission of UK Biobank ©.

In comparison, the 2D configuration yielded a slightly lower but competitive spatial overlap, with a mean DSC of 0.89 SD= 0.005 (median: 0.89). Based on these cross-validation outcomes, the 3D model was selected as the primary architecture for subsequent test dataset benchmarking and population-scale deployment.

4.3 Test performance

The network predictions were benchmarked against the ground truth segmentations on the 24 test images where the imaging specialist broke ties. The model achieved a high spatial overlap, yielding a mean DSC of 0.92 SD= 0.03 (median: 0.93). Boundary distance evaluations via the Hausdorff distance demonstrated tight contour alignment, with a mean boundary separation of 3.58 SD= 2.44 mm (median: 2.72 mm). The mean absolute volume difference between the manual and predicted segmentations was 8249.09 SD= 9289.66 voxels (median: 5739.80 voxels).

4.4 Population-scale results

To demonstrate the scalability and robustness of the developed segmentation network, inference was done at a population scale on the entire cohort of 34,412 subjects from the UK Biobank dataset. Prediction took per-subject inference time (7.3 s) and total compute for the population run (2 Days, 22-hours across the 34,412-subject cohort). Figure 3 shows sample model predictions on unseen data. Figure 1 A illustrates the distribution of the predicted anatomical volumes in milliliters (mL). During initial processing, a subset of cases yielded predicted volumes of zero. A manual review of these statistical outliers revealed that they exclusively represented artifactual imaging errors rather than algorithmic failures. Specifically, these zero-volume predictions corresponded to cases where the

target anatomy was completely omitted from the field of view owing to participant positioning relative to the imaging slabs.

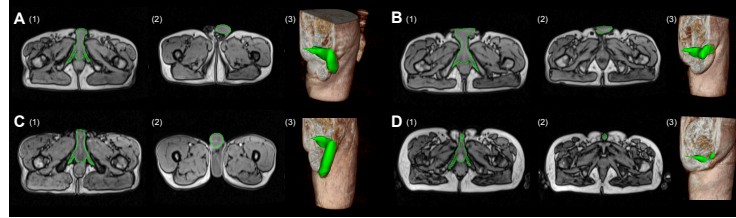


Fig. 3. Examples of nnU-Net predictions. Axial views highlighting the internal (1) and external penile compartments (2), along with corresponding 3D renderings (3), are shown. (B) illustrates likely compression of the external penis due to clothing. (D) depicts a sample with visible internal erectile structures but no externally developed penile shaft. Images reproduced by kind permission of UK Biobank ©.

To evaluate the longitudinal consistency of the network’s predictions, we analyzed a subcohort of $n = 2,282$ subjects who underwent a second, independent imaging session as part of the UK Biobank longitudinal assessment protocol. After excluding participants with zero-volume segmentations in either session ($n = 13$), $n = 2,269$ subjects were retained. Volumetric predictions from the two sessions showed high agreement (Pearson $r = 0.86$; ICC(2,1) = 0.86, 95% CI [0.85–0.87]; Bland–Altman mean bias = 0.81 mL, 95% limits of agreement [−28.4, 30.0] mL; Figure 1 B), indicating good test-retest reliability of the model across independent acquisitions with no substantial systematic drift between sessions.

5 Discussion

This work demonstrates that penile measurements can be converted from a difficult-to-standardize observation into a reproducible, population-scale image-derived phenotype. Its central contribution is the establishment of a volumetric segmentation model capturing both external and internal penile components, the latter being entirely inaccessible to conventional surface anthropometry. On the independent test set, the model reached a Dice of 0.92 against the expert-selected reference, above the 0.82 inter-observer agreement between the two readers. These numbers are not directly comparable: the model is scored against a single selected reference, whereas the inter-observer figure reflects agreement between two readers. A high model-reference Dice therefore indicates close agreement with the convention of the selected reader rather than demonstration of greater anatomical accuracy than human observers. We accordingly interpret the model as performing at expert-annotation level on this benchmark. The comparatively low inter-observer agreement itself reflects genuine anatomical ambiguity

at the proximal root, where expert boundaries diverge (Figure 2). Zero-volume predictions in population deployment were attributable to field-of-view errors, not model failures. Several limitations apply. The training set is modest and the cohort skews older (mean 67 years), which may limit generalizability. All data derive from a single resource (UK Biobank) on a harmonized protocol. Further external validation on independent cohorts is required before broader conclusions can be drawn. The model segments the penis as a single structure; substructure-level delineation (glans, corpora, bulb, crura) would be valuable but is constrained by Dixon MRI resolution. Future work should validate in younger and clinical cohorts and explore sub-organ segmentation as resolution improves.

Prospects of application: This framework introduces penile volume as a scalable image-derived phenotype for population-based male health research. It enables large-scale association studies linking penile anatomy to genetic variants, proteomic and metabolomic profiles, hormonal status, body composition, aging, and reproductive or urogenital traits, thereby advancing research into the biological basis of male reproductive disorders.

Acknowledgments. J.E. was supported by the Medical Scientist Kolleg InFlame funded by the Else Kröner-Fresenius Foundation, Germany.

ASB is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) 464240267 and SFB 1748/1 2026.

This research has been conducted using the UK Biobank Resource under Application Number 100261.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berg, I.v.d., Savenije, M.H., Teunissen, F.R., Pol, S.M.v.d., Rasing, M.J., Melick, H.H.v., Brink, W.M., Boer, J.C.d., Berg, C.A.v.d., Zyp, J.R.v.d.V.v.: Deep learning for automated contouring of neurovascular structures on magnetic resonance imaging for prostate cancer patients. *Physics and Imaging in Radiation Oncology* **26**, 100453 (2023). <https://doi.org/10.1016/j.phro.2023.100453>
2. Cakir, O.O., Pozzi, E., Castiglione, F., Alnajjar, H.M., Salonia, A., Muneer, A.: Penile length measurement: Methodological challenges and recommendations, a systematic review. *The Journal of Sexual Medicine* **18**(3), 433–439 (2021). <https://doi.org/10.1016/j.jsxm.2020.11.012>, <https://doi.org/10.1016/j.jsxm.2020.11.012>
3. Ernsting, J., Beeken, P.N., Ogoniak, L., Kockwelp, J., Roll, W., Hahn, T., Busch, A.S., Risse, B.: Towards population scale testis volume segmentation in DIXON MRI. *Computers in Biology and Medicine* **198**(Pt A), 111139 (2025). <https://doi.org/10.1016/j.compbimed.2025.111139>
4. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R.: 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging* **30**(9), 1323–1341 (2012). <https://doi.org/10.1016/j.mri.2012.05.001>
5. Graf, R., Platzek, P., Riedel, E.O., Ramschütz, C., Starck, S., Möller, H.K., Atad, M., Völzke, H., Bülow, R., Schmidt, C.O., Rüdebusch, J., Jung, M., Reisert, M., Weiss, J., Löffler, M.T., Bamberg, F., Wiestler, B., Paetzold, J.C., Rueckert, D., Kirschke, J.S.: VIBESegmentator: full body MRI segmentation for the NAKO and UK biobank. *European Radiology* **36**(4), 1–15 (2025). <https://doi.org/10.1007/s00330-025-12035-9>
6. Hiort, O., Birnbaum, W., Marshall, L., Wunsch, L., Werner, R., Schröder, T., Döhnert, U., Holterhus, P.M.: Management of disorders of sex development. *Nature Reviews Endocrinology* **10**(9), 520–529 (09 2014). <https://doi.org/10.1038/nrendo.2014.108>, <https://doi.org/10.1038/nrendo.2014.108>
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (02 2021). <https://doi.org/10.1038/s41592-020-01008-z>, <https://doi.org/10.1038/s41592-020-01008-z>
8. Kumar, S., Mehta, P.M., Chaudhary, R.K., Nepal, P., Kumar, D., Tirumani, S.H., Ojili, V.: MR imaging of the penis: What a radiologist needs to know! *Abdominal Radiology* **50**(4), 1790–1810 (2025). <https://doi.org/10.1007/s00261-024-04532-0>
9. Lavdas, I., Glocker, B., Rueckert, D., Taylor, S., Aboagye, E., Rockall, A.: Machine learning in whole-body MRI: experiences and challenges from an applied study using multicentre data. *Clinical Radiology* **74**(5), 346–356 (2019). <https://doi.org/10.1016/j.crad.2019.01.012>
10. Schubert, P., May, M., Höfler, D., Fautz, H.P., Hutter, J., Merten, R., Mansoorian, S., Weissmann, T., Deloch, L., Schonath, M., Belmas, N., Grabenbauer, F., Frey, B., Gaip, U., Axer, B.N., Szkitsak, J., Uder, M., Bert, C., Fietkau, R., Putz, F.: Advancing offline magnetic resonance-guided prostate radiotherapy through dedicated imaging and deep learning-based automatic contouring of targets and neurovascular structures. *Physics and Imaging in Radiation Oncology* **35**, 100825 (2025). <https://doi.org/10.1016/j.phro.2025.100825>
11. Veale, D., Miles, S., Bramley, S., Muir, G., Hodsoll, J.: Am i normal? a systematic review and construction of nomograms for flaccid and erect penis length and

- circumference in up to 15 521 men. *BJU International* **115**(6), 978–986 (2015). <https://doi.org/10.1111/bju.13010>
12. West, J., Leinhard, O.D., 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), e0163332 (2016). <https://doi.org/10.1371/journal.pone.0163332>