Appendix S1: Radiomics feature extraction

In this study, radiomics features quantifying intensity, shape and texture were extracted. Intensity features were extracted using the histogram of all intensity values within the Regions of Interest (ROIs) and included several first order statistics such as the mean, standard deviation and kurtosis. Shape features were extracted by solely using the ROI and included shape descriptions such as the compactness, roundness and circular variance. Additionally, the volume and orientation of the ROI were used. Texture features were extracted using the Gray Level Co-occurrence Matrix, Gray Level Size Zone Matrix Gray Level Run Length Matrix and Neighborhood Grey Tone Difference Matrix. All features were extracted using the defaults for MR images from PyRadiomics.

The used dataset is highly heterogeneous in terms of acquisition protocols. Especially the variations in slice thickness and contrast may cause feature values to be highly dependent on the acquisition protocol. The slice thickness varies between 2.5mm and 10mm. Hence, extracting robust 3D features may be hampered by these variations, especially for the low resolutions. To overcome this issue, all features are extracted per 2D axial slice and aggregated over all slices. Due to the slice thickness and pixel spacing heterogeneity, the images were not resampled. Due to variations in especially the magnetic field strength, echo time, and repetition time, the image contrast highly varies, which will affect the feature values. To overcome this, each 3D MRI is normalized using z-scoring before feature extraction.

The code to extract the features has been published open-source.[1]

[1] https://github.com/MStarmans91/LipoRadiomicsFeatures

Appendix S2: Technical details on decision model creation

The Workflow for Optimal Radiomics Classification (WORC) toolbox[1] makes us of adaptive algorithm optimization to create the optimal performing workflow from a variety of methods. We define a workflow as a sequential combination of algorithms and their respective parameters.

WORC includes algorithms to perform feature imputation, feature selection, feature scaling, oversampling, and machine learning. Feature selection was performed to eliminate features which are not useful to distinguish between WDLPS and lipoma. These included; 1) a group-wise search, in which specific groups of features (i.e. intensity, shape, and the several subgroups of texture features as defined in Supplementary Materials 1) are selected or deleted; 2) a variance threshold, in which features with a low variance are removed; and 3) principal component analysis (PCA), in which only those linear combinations of features were kept which explained a large part of the variance in the features.

Feature scaling was performed to make all features have the same scale, as otherwise the machine learning methods may focus only on those features with large values. This was done through z-scoring, i.e. subtracting the mean value followed by division by the standard deviation. In this way, all features had a mean of zero and a variance of one.

Oversampling was used to make sure the classes (i.e. WDLPS and lipoma) were balanced in the training dataset. These include 1) random oversampling, which randomly repeats patients of the minority class; and 2) SMOTE[2], which creates new synthetic patients using a combination of the patients in the minority class.

Lastly, machine learning methods were used to determine a decision rule to distinguish between WDLPS and lipoma. These included 1) logistic regression; 2) support vector machines; 3) random forests; 4) naive Bayes; and 5) linear and quadratic discriminant analysis.

Most of the included methods require specific settings or parameters to be set, which may have a large impact on the performance. As these parameters have to be determined before executing the workflow, these are so-called "hyperparameters". In WORC, we treat all parameters of all methods as hyperparameters, since they may all influence the decision model creation. Hence, we simultaneously determine which combination of algorithms and hyperparameters performs best.

In the training phase, a total of 100,000 pseudo-randomly generated workflows is created and executed. The workflows are ranked from best to worst based on the F1-score, which is the harmonic average of precision and recall. Due to the large number of workflows executed, there is a chance that the best performing workflow is overfitting, i.e. looking at too much detail or even noise in the training dataset. Hence, to create a more robust model and boost performance, WORC combines the 50 best performing methods into a single decision model, which is known as ensembling. The ensemble is created through averaging of the probabilities, i.e. the chance of a patient being WDLPS or lipoma, of these 50 workflows.

---

[1] Workflow for Optimal Radiomics Classification (WORC). https://github.com/MStarmans91/WORC.

[2] Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In; 2005; Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 878-887.