Appendix 1

Patient data

Segmentation of dorsal striatum (DS) on DaT SPECT images via T1 weighted MRI

Imaging was performed 4±0.5 h following injection of DAT SPECT (^{123}I-ioflupane; 111–185 MBq). Thyroid update was blocked via pre-treatment of subjects with saturated iodine solution (10 drops in water) or perchlorate (1,000 mg) prior to injection. Data acquisition consisted of 128×128 raw SPECT projection data acquired every 3 degrees, 120 projections, 20% symmetric photopeak windows centered on 159 and 122 keV, and a total scan duration of ~30–45 min. A HERMES system (Hermes Medical Solutions, Stockholm, Sweden) was used to perform iterative OSEM reconstruction on the input raw SPECT projection data, for all studies to ensure consistency. Subsequently, PMOD (PMOD Technologies, Zurich, Switzerland) was used for attenuation correction. Ellipses where drawn on the images and Chang 0 attenuation correction were applied invoking a site-specific mu as empirically derived from phantom data (as acquired in site initiation for the trial). Following this, standard 3D Gaussian post-smoothing (6.0 mm FWHM) was applied (43,46).

As shown in *Figure S1*, pre-processing process is included 3 parts. We firstly resized pixels of images as 1×1×1. After fixing orientation of images, we modified intensity bias of images. The original aspect ratio of the MR images was maintained. After pre-processing section, we applied these images to Free surfer for segmenting region of interests (ROI) such as left and right caudate as well as putamen. Co-registration of SPECT on MRI were performed through two steps. In first step, we manually co-reiterated SPECTS on their MRI through Mango software. We tried various combinations of transform options, cost function options and search cost options on each image. We experimentally considered best solution for each image. There were various options for transform options includes (I) 2D rigid body, (II) translation only, (III) rigid body, (IV) global scale, (V) full scale, and (VI) full affine. There were also different options for cost function options and search cost options included: (I) correlation ratio, (II) mutual information, (III) norm mutual information, (IV) normalized correlation, and least square. Finally, we employed the rigid co-registration algorithm accompanied with normalized mutual information on the co-registered DAT SPECT images as well as Gaussian smoothing kernel with a width of 7 mm for all images. In the end, we overlaid the structures segmented in MRI on the co-registered SPECT images. After co-registration stage, we applied each mentioned ROI to the SERA to extract radiomics features. SERA has been extensively standardized in reference to the Image Biomarker Standardization Initiative (ISBI) (112), and studied in multi-center radiomics standardization publications by the IBSI (113) and the quantitative imaging network (QIN) (114). There is a total of 487 standardized radiomics features in SERA, including: 79 first-order features (morphology, statistical, histogram and intensity-histogram features), 272 higher-order 2D features, and 136 3D features. We included all 79 first-order features and 136 3D features (54,113,115).

List of features used

We considered the following 981 features (as shared publicly): multiple Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) measures, a range of task/exam performances, socioeconomic/family histories, genetic features, and SPECT image features. Segmentation of regions-of-interest (ROIs; left and right

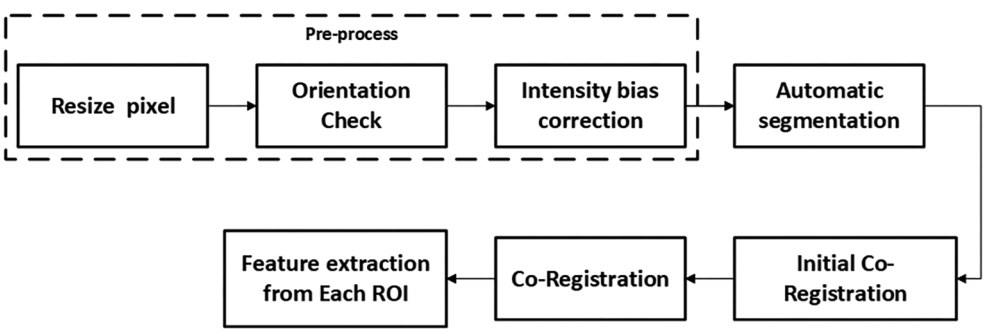

Figure S1 Diagram of image processing and radiomics feature extraction based on MRI images.

https://dx.doi.org/10.21037/qims-21-425

both caudate and putamen) on DaT SPECT images were performed via MRI images. Radiomic features (RFs) were extracted for each ROI using our standardized SERA software. For consistency, we only included patients who were off medication (e.g., Levodopa/dopamine agonist) for >6 hours prior to testing/imaging (47). we separately collected information for patients based on each year. Subsequently, timeless datasets were constructed by appending cross sectional datasets within a single set of data. This approach aims to gather data with larger number of subjects and features.

Machine learning algorithms

Feature extraction algorithms

Principal component analysis (PCA)

PCA is a known tool for linear dimensionality reduction and feature extraction. Using an orthogonal transformation enables us to convert a dataset with correlated variables into new dataset with linearly uncorrelated variables called principal components. The first principle component has the highest variance in compered to other principle components. Moreover, number of principal components is less than or equal to the number of original variables (58).

Kernel PCA

Kernel PCA is an extended nonlinear form of the PCA using techniques of kernel methods. it is more useful to extract the complicated spatial structure of high-dimensional features in compared to simple PCA. Thus, the Kernel PCA is increasingly using in machine learning application (59).

t-distributed stochastic neighbor embedding (t-SNE)

t-SNE is a machine learning algorithm which performs non-linear dimensionality reduction to embed high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high dimensional object by a two or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability (60).

Factor analysis (FA)

FA, as an analytic technique enables us to reduce a large number of correlated variables to a smaller number of dimensions. The goal of the FA is to achieve parsimony by minimizing explanatory concepts to explain the maximum amount of common variance in a correlation matrix (61).

Sammon mapping algorithm (SMA)

SMA is a non-linear algorithm which maps a high dimensional dataset to a low dimensional dataset. It also preserves the structure of inter-point distances in the high dimensional dataset in the lower dimensional space.it works based on minimizing error function (called Sommon error or Stress error), as shown in equation [1] (62).

$$E = \frac{1}{\sum_{i<j} d_{ij}^*} \sum_{i<j} \frac{\left(d_{ij}^* - d_{ij}\right)^2}{d_{ij}^*} \quad [1]$$

Where d_{ij}* is the distance between ith and jth datapoint in the original space and d_{ij} is the distance between their projections.

The stress function was improved using left Bregman divergence and right Bregman divergence (63).

Isomap algorithm (IsoA)

IsoA, as a nonlinear dimensionality reduction algorithm, is one of low-dimensional embedding methods. It was employed for compute a quasi-isometric in order to embed low-dimensional points of a set of high-dimensional data points (64).

LandMark Isomap algorithm (LMIsoA)

LMIsoA is a fast IsoA which is faster than the Isomap. It works based on landmark MDS (Multi-Dimension Scaling) so that it selects a group of points termed as Landmarks and implements classical MDS on them. After computing the shortest path from each data point to the landmark points, the geodesic distance matrix is applied on classical MDS to find the low-dimensional embedding of the landmark points (65).

Laplacian eigenmaps algorithm (LEA)

LEA, as a non-linear dimensionality reduction, aims to build a graph from neighborhood connections of the dataset. The discrete approximation of the low-dimensional manifold in the high-dimensional space are considered as connections between nodes which constructing by each data point. Minimizing the cost function based on the graph enables us to guarantee that close points are mapped close to each other in the low-dimensional space (66,67).

Locally linear embedding algorithm (LLEA)

LLEA, as an unsupervised learning algorithm, computes

https://dx.doi.org/10.21037/qims-21-425

low dimensional, neighborhood preserving embeddings of high dimensional data. It works based on exploiting the local symmetries of linear reconstructions in order to find nonlinear structure in high dimensional data (68).

Multidimensional scaling algorithm (MDSA)

MDSA, known as Principal Coordinates Analysis, provides a visual representation of the pattern of proximities (i.e., correlation matrix or distance matrix) among a set of objects. It aims to map these dissimilarities as distances between points in a low dimensional space so that these distances correspond as closely as possible to the dissimilarities (69).

Diffusion map algorithm (DMA)

DMA, as a non-linear technique, reduces high dimensional space to low dimensional space by re-organising data according to parameters of its underlying geometry. It computes a family of embeddings of dataset into Euclidean space using coordinates from the eigenvectors and eigenvalues of a diffusion operator on the data. These distances among datapoints in the embedded spaces are equal to diffusion distances between probability distributions centered at those points (70,71).

Stochastic proximity embedding algorithm (SPEA)

SPEA, as a novel self-organizing algorithm, produces meaningful underlying dimensions from proximity data. It reduces high dimensional dataset to low-dimensional Euclidean embeddings so that the similarities between a set of related observations preserve. This algorithm initially selects a random configuration and then adjusting their coordinates according to iteratively refining it by repeatedly selecting pairs of objects at random (72).

Gaussian process latent variable model (GPLVM)

GPLVM, as a dimensionality reduction method, is a flexible Bayesian non-parametric modeling method that learns a low-dimensional representation of high-dimensional data by a Gaussian process. In this case, the kernel and learning hyperparameters of Gaussian process regression are selected to describe the mapping of high dimensional dataset to low dimensional dataset (73,74).

Stochastic neighbor embedding algorithm (SNEA)

SNE, as a dimensionality reduction method, is a probabilistic method of embedding objects. It maps a high-dimensional vectors or pair wise dissimilarities into a lower dimensional space so that neighbor identities are preserved. Dissimilarities, providing via applying a Gaussian model to each object in the high-dimensional space, are used to define a probability distribution which has potential neighbors of the object (75).

Symmetric stochastic neighbor embedding algorithm (Sym_SNEA)

SNEA has slow convergence; a fast SNE algorithm was proposed that is approximately 4-6 times faster. This algorithm works based on a trust-region method to discover a reliable direction as well as efficient step size with the help of a quadratic model of the objective function (76).

Autoencoders algorithms (AA)

AA, a three-layered neural network, constructs the "building block" of deep learning. It can be converted high-dimensional data to low-dimensional codes by training a multilayer neural network. Gradient descent can be employed for updating weight matrix in the AA (77).

K-Means algorithm (KMA) for unsupervised clustering

The KMA, as a type of unsupervised learning method, tries to partition the dataset into K distinct non-overlapping sub-groups. It works based on minimizing the sum of the squared distance between the subjects and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) similar subjects in a cluster (116,117).

We considered 3 first features of the first subject as initial centroid values to cluster patients (original sub-cluster), while we used 3 first columns of the first patient as initial centroids to identify progression trajectories. We also discovered that different combination of initial centroids resulted in same sub-clusters and merely the order of sub-clusters was changed. In future work, we plan to explore other initialization methods. Moreover, we considered 1,000 epochs (maximum iteration) for this algorithm.

Clustering evaluation methods

Calinski-Harabatz criterion

The Calinski-Harabasz index of a clustering is the ratio of the between-cluster variance (which is essentially the variance of all the cluster centroids from the dataset's grand centroid) to the total within-cluster variance. The within-cluster variance will decrease as k increases; the rate of decrease should slow down past the optimal k. The between-cluster variance will increase as k, but the rate of increase should slow down past the optimal k. So, in theory,

https://dx.doi.org/10.21037/qims-21-425

the ratio of between-cluster variance to within-cluster variance should be maximized at the optimal k. The ratio is formally defined as:

$$C(k)=\frac{Trace(B)}{Trace(W)}\frac{n-k}{k-1} \quad [2]$$

where B is the between-cluster covariance matrix [so high values of Trace(B) denote well-separated clusters], W is the within-cluster covariance matrix [so low values of Trace(W) correspond to compact clusters], n is the number of the data points and k is the number of the clusters (82).

Bayesian information criterion (BIC)

In statistics, the BIC is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC). Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. The BIC is formally defined as (83):

$$BIC = k\ln(n)-2\ln(\hat{L}) \quad [3]$$

where

$\hat{L}$ = the maximized value of the likelihood function of the model

x = the observed data;

n = the number of data points in x, the number of observations, or equivalently, the sample size;

k = the number of parameters estimated by the model.

Elbow criterion

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use (81).

Classifiers

Decision tree classification (DTC)

DTC technique is one of the most popular techniques in the emerging field of data mining. There are various methods for constructing the DTC. Induced decision tree (ID3) is the basic algorithm for constructing the DTC (84). There are many algorithms based on classification that is sample based, neural networks, Bayesian networks, support vector machine, and decision tree. The DTC classifies samples by sorting them down the tree from the root to some leaf node, which provides the classification of samples. Each node in the tree specifies a test of some attribute of the sample and each branch descending from that node corresponds to one of the possible values for this attribute (104). In this specific work, the maximum depth was not set so the algorithm would continue until all leaves were pure. To measure the quality of a split, a "Gini" function was used (GINI function describes the impurity of each node; each child node was purer than its parent node so that the GINI function was minimized).

Library for support vector machines (Lib_SVM)

LIBSVM is a library for support vector machines (SVM) (87). SVM, as a supervised learner, was initially designed for binary classification, defined by a separating hyperplane. Optimal hyperplane, which categorizes new examples, is regulated by labeled training data. In two-dimensional space, this hyperplane is a line dividing a plane in two parts where in each class lay in either side (88). To extend SVM to the multi-class scenario, several classification models were proposed such as the one by Crammer and Singer (89). It was replaced the misclassification error of an example with the piecewise linear bound.

K nearest neighborhood classifier (KNNC)

k-NN, as a supervised and non-parametric algorithm, employs for classification and regression tasks. On the other words, it works based on instance-based learning, where the function is only approximated locally. In both cases, the input consists of the k closest training examples in the feature space as well as output is a class membership. Thereby, objects are classified by a plurality vote of their neighbors (90,91).

Ensemble leaner classifier (ELC)

ELC, as a supervised learner, works according to voting process of multiple classifiers. All classifiers, combined to solve a common problem, participate in prediction process. It mostly results in better predictive performance than use of a sole classifier (92,93).

Linear discriminant analysis classifier (LDAC)

LDAC, as a generalization of Fisher's linear discriminant, is employs in machine learning area to discover a linear combination of features that separates two or more classes of objects. Thus, the combination may be used as a linear classifier or as a dimensionality reduction before the

https://dx.doi.org/10.21037/qims-21-425

classification (94,95).

New probabilistic neural network classifier (NPNNC)

PNNC, as a supervised feed-forward neural network with a complex structure consists of an input layer, a pattern layer, a summation layer and an output layer. A single training parameter (probability density functions) is considered to activate the neurons in the pattern layer (96,97).

Error-correcting output codes model classifier (ECOCMC)

ECOCMC, as a general classification framework and a supervised algorithm, needs multiple classifiers, similarly to the ELC. It works based on two stages: encoding and decoding. Encoding phase consists of designing coding matrix. The columns and rows of the matrix show binary classifiers and codewords for classes respectively. Designing of a coding matrix can be made using binary coding and ternary coding. In decoding phase, we need to find which one of the classes' codewords is the closest one the test example's codeword (98,99).

Multi-layer perceptron-back propagation (MLP-BP)

A multilayer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output so it is a modified MLP that uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that are not linearly separable, or separable by a hyper plane (46,100). In this specific work, we used a three-layer neural network and the number of neurons in each layer was adjusted via automated machine learning hyperparameter tuning automatically.

Random forest algorithm (RFA)

Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them (102,20). Depth of structure was adjusted via automated machine learning hyperparameter tuning automatically. Number of trees and number of splits were set to 1000 and 5, respectively.

Recurrent neural network (RNN)

Recurrent neural network is a deep learning algorithm. The RNN as fundamentally different neural network from feed-forward architectures was investigated for modelling of nonlinear behavior (41,104). In this work, we used a model with many inputs to one output. In this specific work, we used a three-layer neural network and the number of neurons in each layer was adjusted via automated machine learning hyperparameter tuning automatically.

Automated machine learning hyperparameter tuning

In this work, automated machine learning hyperparameter tuning was employed to automatically adjust intrinsic parameters such as the number of neurons, and the number of layers in the classification algorithms. We applied this approach to various algorithms such as LOLIMOT, RBF, RNN, MLP-BP, RFA to automatically tune the parameters. Automated tuning, which was implemented with our in-house code, executes an error minimization search scheme to optimize the hyperparameters starting with the random initialization. Employing this approach enables us to pursue a systematic trial-and-error search scheme for tuning the parameters (38).

Hoteling's t squared test

This statistical test is used to evaluate the equality of the mean vectors of two populations (with n1 and n2 samples). Each of two groups has p features. Assume that population 1 is distributed as $N_p(\mu_1, \Sigma_1)$ and population 2 is distributed as $N_p(\mu_2, \Sigma_2)$, where $N_p(\mu,\Sigma)$ is the p-variable multivariate normal distribution with mean vector μ and covariance matrix Σ. The null hypothesis that $\mu_1 = \mu_2$ can be tested using the test statistic:

$$T2 = \frac{n_1 n_2}{n_1 + n_2}\left(Y_1 - Y_2\right)' S_{pooled}^{-1}\left(Y_1 - Y_2\right) \quad [4]$$

where Y_1 and Y_2 are the two sample mean vectors, n_1 and n_2 are the two sample sizes, and S_{pooled}^{-1} is the inverse of the pooled covariance matrix which is calculated using:

$$S_{Pooled} = \frac{\left(n_1 - 1\right)S_1 + \left(n_2 - 1\right)S_2}{n_1 + n_2 - 2} \quad [5]$$

Here, S_1 and S_2 are the estimated covariance matrices calculated from the two samples. If we make the additional assumption that $\Sigma_1 = \Sigma_2$, T2 follows Hotelling's T-squared distribution when the null hypothesis is true. That is, $T2 = T^2_{P,n_1+n_2-2}$. Reject the null hypothesis if $T2 \geq T^2_{P,n_1+n_2-2}$. Note that rejecting the null hypothesis concludes that at

https://dx.doi.org/10.21037/qims-21-425

least one pair of the p sets of group response means are unequal (105).

References

112. Image biomarker standardisation initiative. Available online: https://arxiv.org/abs/1612.07003
113. Hall JF, Crocker TF, Clarke DJ, Forster A. Supporting carers of stroke survivors to reduce carer burden: development of the Preparing is Caring intervention using Intervention Mapping. BMC Public Health 2019;19:1408.
114. McNitt-Gray M, Napel S, Jaggi A, Mattonen SA, Hadjiiski L, Muzi M, et al. Standardization in Quantitative Imaging: A Multicenter Comparison of Radiomic Features from Different Software Packages on Digital Reference Objects and Patient Data Sets. Tomography 2020;6:118-28.
115. Ashrafinia S, Dalaie P, Yan R, Huang P, Pomper M, Schindler T, Rahmim A. Application of Texture and Radiomics Analysis to Clinical Myocardial Perfusion SPECT Imaging. J Nucl Med 2018;59(Supplement 1):94.
116. Kanungo T, Member S, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 2002;24:881-92.
117. Francis BK, Babu SS. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. J Med Syst 2019;43:162.

https://dx.doi.org/10.21037/qims-21-425