Supplemental Data:

## Additional methods

Random forests were used as they have been shown to be very successful for a range of different classification problems, are sufficiently robust to noisy data, and can be used for identifying the most salient features (i.e., predictor importance of different Sniffin’ sticks). A classification random forest is an ensemble method that constructs several classification trees, whereby each classification tree provides a classification (e.g., good smell or poor smell) using input features (answers to each individual Sniffin’ stick), and the most common prediction across different trees is used as the final prediction from the random forest. We employ random forests for binary classification in this study, which is aimed at differentiating poor versus good smell. Future studies could extend this framework to identify granular variations in smell status using a multi-class classifier.

Though the optimal choice of cross validation strategy within the field of machine learning remains contentious,3 one view is that LOSO CV more closely reflects the clinical scenario in that it allows for the accuracy of predictions for a previously unseen individual to be determined, based on knowledge gleaned from a larger, mutually exclusive training dataset. Thus, we used LOSO CV in this study.

In this study, a single random forest model comprised 500 classification trees, and each of the 10 balanced groups included 534 participants. For a given balanced group, using the LOSO CV scheme, we trained the random forest model using data from 533 participants, and tested the prediction on the remaining 1 participant. For each balanced group, the process of training and testing was repeated a total of 534 times, so that every participant ended up in the testing data exactly once (resulting in 534 random forest models, each comprising 500 tress). The process of generating predictions for 534 participants was repeated for a total of 10 balanced groups. The 5340 out-of-sample predictions were collated and used to calculate the AUCs shown in the first row (“Overall PD poor smell model accuracy”) of Table 2.

Table e-1 | Sniffin’ stick combinations proposed for A) the detection of poor smell B) distinguishing individuals with PD from controls

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Total sticks | Orange | Leather | Cinnamon | Peppermint | Banana | Lemon | Licorice | Turpentine | Garlic | Coffee | Apple | Clove | Pineapple | Rose | Anise | Fish |
| Detecting poor smell | Discovery | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Q StickHummel et al.1 | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Brief SniffinMueller et al.2 | 5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Distinguishing PD versus controls | Boesveldt et al.3 | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Casjens et al.4 | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Mahlknecht et al.5  | 8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Table e-2 | Sniffin’ stick identification in individuals with Parkinson’s and controls, with poor and good smell. aComprising individuals with hyposmia or functional anosmia bcomprising individuals with normosmia or super smell cp value determined using a chi squared test

|  |  |  |
| --- | --- | --- |
|  | Poor smella | Good smellb |
| Percentage of individuals who identified each stick correctly |  | Percentage of individuals who identified each stick correctly |  |
| Stick number | Stick smell | Controls | PD | pc | Controls | PD | pc |
| 1 | Orange | 76 | 63 | 0.06 | 96 | 91 | 0.02 |
| 2 | Leather | 48 | 41 | 0.37 | 84 | 76 | 0.06 |
| 3 | Cinnamon | 43 | 33 | 0.15 | 66 | 68 | 0.59 |
| 4 | Peppermint | 83 | 58 | <0.01 | 97 | 90 | <0.001 |
| 5 | Banana | 67 | 39 | <0.001 | 92 | 78 | <0.0001 |
| 6 | Lemon | 30 | 22 | 0.20 | 61 | 56 | 0.23 |
| 7 | Licorice | 54 | 26 | <0.0001 | 88 | 72 | <0.0001 |
| 8 | Turpentine | 33 | 32 | 0.92 | 47 | 43 | 0.41 |
| 9 | Garlic | 72 | 54 | 0.02 | 84 | 83 | 0.77 |
| 10 | Coffee | 43 | 29 | 0.04 | 78 | 67 | 0.01 |
| 11 | Apple | 9 | 13 | 0.38 | 33 | 31 | 0.63 |
| 12 | Clove | 65 | 39 | <0.001 | 88 | 76 | <0.01 |
| 13 | Pineapple | 33 | 34 | 0.89 | 76 | 66 | 0.01 |
| 14 | Rose | 54 | 48 | 0.42 | 85 | 67 | <0.0001 |
| 15 | Anise | 54 | 22 | <0.0001 | 94 | 74 | <0.0001 |
| 16 | Fish | 76 | 56 | <0.01 | 99 | 94 | 0.01 |

Table e-3 | Area under the curve values for MLA-trained 3 Sniffin’ stick models in the detection of hyposmia or functional anosmia by age and sex. As calculated across 5340 trained models where the data used to train models and the data used to assess model accuracy were mutually exclusive. a Given the presence of a single class (with all of the individuals within the group having a good sense of smell) it was not possible to calculate AUC values for the 21 to 30 age group.

|  |  |
| --- | --- |
| Overall PD poor smell model accuracy | Sex |
| OVERALL | MALE | FEMALE |
| Age groupa | OVERALL |  | 0.96 (0.95-0.96) | 0.95 (0.93-0.96) |
| 31 to 40 | 0.82 (0.67-0.92) | 0.80 (0.56-0.95) | 0.83 (0.58-0.96) |
| 41 to 50 | 0.95 (0.92-0.97) | 0.95 (0.89-0.98) | 0.95 (0.91-0.98) |
| 51 to 60 | 0.94 (0.92-0.95) | 0.96 (0.95-0.97) | 0.87 (0.83-0.91) |
| 61 to 70 | 0.98 (0.97-0.98) | 0.98 (0.97-0.99) | 0.98 (0.96-0.99) |
| 71 to 80 | 0.94 (0.93-0.95) | 0.92 (0.90-0.94) | 0.98 (0.96-0.98) |
| 81 and over | 0.96 (0.92-0.98) | 0.94 (0.89-0.97) | 1.00 (1.00-1.00) |

Figure e-1 | Histograms demonstrating the distribution of AUC values for the detection of poor smell in individuals with RBD from the Discovery study, when using the answers to all possible 3 Sniffin’ stick combinations from participants with A) PD poor smell/control good smell, B) Control poor smell/control good smell and C) PD poor smell/control good smell to train MLAs.

A significant difference between the three distributions exists (p<0.0001) as determined by two-sample pairwise Kolmogorov-Smirnov tests.



# References

1. Hummel T, Pfetzing U, Lotsch J. A short olfactory test based on the identification of three odors. Journal of neurology 2010;257:1316-1321.

2. Mueller C, Renner B. A new procedure for the short screening of olfactory function using five items from the "Sniffin' Sticks" identification test kit. Am J Rhinol 2006;20:113-116.

3. Boesveldt S, Verbaan D, Knol DL, et al. A comparative study of odor identification and odor discrimination deficits in Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2008;23:1984-1990.

4. Casjens S, Eckert A, Woitalla D, et al. Diagnostic value of the impairment of olfaction in Parkinson's disease. PloS one 2013;8:e64735.

5. Mahlknecht P, Pechlaner R, Boesveldt S, et al. Optimizing odor identification testing as quick and accurate diagnostic tool for Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2016;31:1408-1413.