Olfactory testing in Parkinson's disease & REM behavior disorder; a machine learning approach

Christine Lo, BMedSci (Hons), MRCP[1,2]; Siddharth Arora, BTech, DPhil[1,3]; Yoav Ben-Shlomo, PhD, MRCP[4], Thomas R. Barber, MA, MBBS, MRCP[1,2]; Michael Lawton, PhD[4]; Johannes C. Klein, MD, PhD[1,2]; Sofia Kanavou, MSc[4]; Annette Janzen, MD[5]; Elisabeth Sittig, SN[5]; Wolfgang H. Oertel, MD[5,6]; Donald Grosset, MD[7]; Michele T. Hu, PhD, FRCP[1,2].

[1]Oxford Parkinson's Disease Centre (OPDC), University of Oxford, UK

[2]Nuffield Department of Clinical Neurosciences, University of Oxford, UK

[3]Saïd Business School, University of Oxford, Oxford, UK

[4]Population Health Sciences, University of Bristol, Bristol, UK

[5]Department of Neurology, Philipps University Marburg, Marburg, Germany

[6]Institute for Neurogenomics, München Helmholtz Center for Health and Environment, Neuherberg München, Germany

[7]Institute of Neurological Sciences, Queen Elizabeth University Hospital, Glasgow, UK

Corresponding author:

Name: Dr Christine Lo

Address: Nuffield Department of Clinical Neuroscience, Level 6, West Wing, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK

Telephone: +44 (0) 1865 223166

Email: Christine.lo@nhs.net

Supplemental Data: Supplementary Figures: 1

Supplementary Tables: 3

Title character count: 93

Abstract: 247

Paper word count: 4433

References: 50

Figures: 4

Tables: 3

Study funding: This study was funded by the Monument Trust Discovery Award from Parkinson's UK (J-1403) and supported by the Oxford National Institute for Health Research (NIHR) Biomedical Research Center (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Search terms: [16] Clinical neurology examination, [54] Cohort studies, [248] Parasomnias, [165] Parkinson's disease/Parkinsonism, Olfaction

Abstract

Objective

We sought to identify an abbreviated test of impaired olfaction, amenable for use in busy clinical environments in prodromal (isolated REM sleep Behavior Disorder (iRBD)) and manifest Parkinson's disease (PD).

Methods

890 PD and 313 control participants in the Discovery cohort study underwent Sniffin' stick odour identification assessment. Random forests were initially trained to distinguish individuals with poor (functional anosmia/hyposmia) and good (normosmia/super-smeller) smell ability using all 16 Sniffin' sticks. Models were retrained using the top 3 sticks ranked by order of predictor importance. One randomly selected 3-stick model was tested in a second independent PD dataset (n=452) and in two iRBD datasets (Discovery n=241; Marburg n=37) before being compared to previously described abbreviated Sniffin' stick combinations.

Results

In differentiating poor from good smell ability, the overall area under the curve (AUC) value associated with the top 3 sticks (Anise/Licorice/Banana) was 0.95 in the development dataset (sensitivity:90%, specificity:92%, positive predictive value:92%, negative predictive value:90%). Internal and external validation confirmed AUCs≥0.90. The combination of 3-stick model determined poor smell and an RBD screening questionnaire score of ≥5, separated iRBD from controls with a sensitivity, specificity, PPV and NPV of 65%, 100%, 100% and 30%.

Conclusions

Our 3-Sniffin'-stick model holds potential utility as a brief screening test in the stratification of individuals with PD and iRBD according to olfactory dysfunction.

Classification of Evidence

This study provides Class III evidence that a 3-Sniffin'-stick model distinguishes individuals with poor and good smell ability and can be used to screen for individuals with iRBD.

Introduction

Paragraph 1

Olfactory dysfunction is evident in up to 90% of individuals with early Parkinson's disease (PD)[1, 2] and demonstrates concordance with dopaminergic deficit.[3, 4] PD stratification according to baseline poor sense of smell, predicts individuals at greater risk of accelerated cognitive decline and dementia.[5-10]

Paragraph 2

The tendency of hyposmia to predate motoric PD by up to twenty years has led to its proposition as a prodromal marker.[11] Amongst non-motor markers, abnormal olfaction (hazard ratio of 2.62) most strongly predicts disease conversion in isolated REM sleep behavior disorder (iRBD);[12] an additional age cut-off of ≥55 suggested to further improve selection for future neuroprotective trials.[13]

Paragraph 3

Rates of subjectively reported and objectively tested hyposmia can differ by upwards of 20%.[14] The Sniffin' sticks test[15] and the UPSIT[16] are two popular tests of olfaction. The former benefits from relative cost effectiveness due to multi-user presentation of felt-tip style pens; the latter involves single-use scratch cards. The Sniffin' sticks test has been extensively studied and validated across different populations.[17-20]

Paragraph 4

The existence of a PD-specific pattern of smell loss remains contentious[21, 22] yet the need for an abbreviated smell test, is well acknowledged.[23-28] Here, our aim was to derive an abbreviated Sniffin' stick test to identify individuals with a poor sense of smell, rather than using Sniffin' stick answers to distinguish disease groups, as in previous studies.[21, 23, 24, 26, 28-

[32] Our test is validated using one independent PD and two independent iRBD datasets, and its combination with the RBDSQ[33] in screening for iRBD is explored.

Methods

Paragraph 5

Primary research questions

- Can an abbreviated Sniffin' stick test identify individuals with a poor sense of smell? (Class III evidence)
- Can an abbreviated Sniffin' stick test be used to screen for individuals with iRBD? (Class III evidence)

Participants

A) Development dataset

The development dataset comprised data collected from 890 participants with idiopathic Parkinson's and 313 Controls (age matched, without a personal or family history of PD or a related condition; the spouses and friends of participants with PD) who were enrolled in the longitudinal Oxford Discovery Cohort Study as previously described.[2, 34]

B) Three mutually exclusive, independent validation datasets comprised:

Paragraph 6

1. 241 participants with iRBD enrolled in the Oxford Discovery Cohort Study
2. 452 participants with Parkinson's in the Tracking cohort study
3. 37 participants with iRBD recruited by the Department of Neurology, University of Marburg, Germany

C) A longitudinal control dataset where data was treated as independent of baseline data collected from the same Control participants who contributed data to the development dataset (n=40)

The Oxford Discovery and Tracking PD cohorts are both based within the UK and recruited non-overlapping participants with Parkinson's, who fulfilled the United Kingdom PD Brain

Bank criteria for probable PD, within 3.5 years of diagnosis.[35, 36] The two cohorts share many similarities and have been used to validate each other's findings.[36, 37] Inclusion of individuals with Parkinson's was contingent upon trained researchers attributing a probability of PD of at least 90% at the latest clinic visit. All participants with iRBD had had a polysomnogram confirming their clinically suspected diagnosis of iRBD, in line with International Classification of Sleep Disorders criteria.[38]

Paragraph 7

Standard protocol approvals, registrations, and patient consents

Studies were prospectively approved by the local research ethics committee and all participants provided written informed consent prior to any study-related procedures.

Paragraph 8

Assessments

All participants were seen face to face in clinic at baseline. At each in-person clinic visit, smell was assessed using the 16 Sniffin' sticks odor identification test (Burghart Instruments, Wedel, Germany) where 16 felt-tip pens, each containing an odor, were presented in turn by being held 2cm centred in front of both nostrils, with participants choosing from one of four options provided.[15, 39] Sniffin' sticks were stored at room temperature, out of direct sunlight, in accordance with manufacturer instructions, and their replacement directed by the best before date displayed on each stick.

Definitions

Normative data from 9139 healthy individuals, were used to define age- and sex-specific percentiles for the total number of Sniffin' sticks correctly identified.[39] Applying previously described classification criteria, functional anosmia, a somewhat discordant entity, was the

Paragraph 9

label attached to a score of 8 sticks or fewer and represented the limit 90% of patients with anosmia would correctly identify.[40] Individuals not already classified as having functional anosmia, but with total Sniffin' stick scores below the 10th percentile for their age and sex, were classified as having hyposmia. Similarly, those with scores above the 90th percentile were classified as having super smell and the remaining individuals were classified as normosmic.[39]

Paragraph 10

Analyses

Analyses were performed using MATLAB® software (R2018a; Mathworks®, USA). Data analysed from the Oxford Discovery Study were collected between 27th September 2010 and 28th May 2019. Only complete sets of data were analysed; incomplete data pertaining to 7/320 (2%) of Controls, 14/255 (5%) of individuals with RBD and 39/929 (4%) of individuals with PD were excluded from analysis (Figure 1).

Paragraph 11

Developing the PD poor smell model using the Development dataset

With the aim of developing a model capable of identifying individuals with a poor sense of smell, independent of disease aetiology, data from individuals at extremes was utilised (see *OBJECTIVE 6* for the effect of different data combinations). A PD poor smell group was formed from participants with Parkinson's and hyposmia or functional anosmia (n=721), and a control good smell group was formed from controls with normosmia or super smell (n=267). Only baseline data were used to ensure that each participant contributed a single set of data, thus ensuring the equal importance of each participant during modelling.

There was a discrepancy between the number of individuals in the PD poor smell group and the control good smell group, resulting in an imbalanced dataset, and leading to the potential for a higher priority to be given to the majority class compared to the minority class during model training.[41] Prior to training machine learning algorithms (MLA), the data were thus balanced by randomly under-sampling the majority class.[42] Specifically, a number of participants equal to that in the control good smell group were randomly selected from the PD poor smell group. Having formed a balanced group with a 1:1 ratio of PD poor smell to control good smell, leave one subject out (LOSO) cross validation was performed to assess the generalizability of the trained model to previously unseen data. Random forests, a commonly used machine learning algorithm,[43] were trained using all of the data in the balanced group with the exception of the data from one individual on whom the accuracy of the trained model was tested. The process was repeated for each individual within the balanced group, resulting in a total of 534 predictions, i.e. one prediction (PD poor smell or control good smell) for each participant that belonged to the validation data. The above process was further repeated for 10 different randomly formed balanced groups in total; resulting in a total of 5340 models, which were then used to calculate the area under the curve (AUC) values. Further methodological details are available from Dryad (Additional Methods).

Paragraph 12

For improved clarity, we detail below our key research objectives and the methodological approaches employed to address each objective. Results are subsequently reported using the same numbering system, i.e. Result 1 (Results section) refers to Objective 1 (Methods section).

Paragraph 13

OBJECTIVE 1: To determine the relative importance of individual Sniffin' sticks in the identification of poor smell in the Development dataset

Paragraph 14

Using the individual answers to all 16 sticks, each trained random forest classifier comprised 500 trees. The predictor importance of each stick (as derived from the random forest algorithm using Gini's diversity index criteria for binary splitting) was averaged across all the trained models in order to rank the sticks in descending order of importance.

OBJECTIVE 2: To assess the classification accuracy associated with different numbers of Sniffin' sticks used in the Development dataset

Paragraph 15

5340 models were further trained across another 10 different randomly formed balanced groups for each incremental number of Sniffin' sticks, cumulated in order of descending MLA-identified predictor importance. AUC values were calculated using only the validation sample in order to evaluate overall model accuracy.

Paragraph 16

In creating an abbreviated smell test, it was necessary to reach a compromise between the number of sticks administered as part of the test and the associated AUC value. A minimal number of 3 sticks was selected to comprise the abbreviated test on the basis that the improvement in AUC from the addition of two sticks to that of a single stick, was more than three times the improvement in AUC associated with the cumulative addition of all other 13 sticks. A 3-stick model also permitted a direct comparison with other previously described 3-stick abbreviated smell tests.

OBJECTIVE 3: To validate the use of a 3-Sniffin'-stick model for the detection of poor smell using three independent PD and iRBD validation datasets

Paragraph 17

One 3-stick PD poor smell/control good smell model, henceforth referred to as the 3-Sniffin'-stick model, was randomly selected out of the 5340 models trained using the same top 3 sticks and the Development dataset. Providing internal validation, its accuracy in identifying poor smell/good smell was assessed using baseline olfactory testing data from participants with iRBD in the Oxford Discovery cohort. The accuracy of the 3-Sniffin'-stick model was furthermore externally evaluated in the independent Tracking PD cohort and in a second independent iRBD cohort (Marburg).

Paragraph 18

OBJECTIVE 4: To compare the accuracy of the 3-Sniffin'-stick model in the detection of poor smell with other previously published stick combinations in the Development dataset

The accuracy of the 3-Sniffin'-stick model, using the 3 sticks identified through random forests as having the highest predictor importance for the detection of poor smell, was compared with: A) the 3 sticks previously proposed to constitute the Q stick test,[27] B) the Brief Sniffin' Stick test previously investigated as a screening test for poor smell by Mueller et al.[25] and C) stick combinations suggested to distinguish individuals with PD from controls by Casjens et al.,[26] Boesveldt et al.[29] and Mahlknecht et al.[24] (Data available from Dryad (Table e-1)).

Paragraph 19

OBJECTIVE 5: To compare the accuracy of the 3-Sniffin'-stick model in the detection of poor smell with all other possible 3-stick combinations using a composite validation dataset comprising Discovery iRBD, Marburg iRBD and Tracking PD datasets

There are a total of 560 unique 3-stick combinations of the 16 Sniffin' sticks. To compare their accuracy at detecting poor smell, using baseline data from the Development dataset, a

single balanced group was formed comprising an equal number of randomly selected individuals from the PD poor smell group and the control good smell group. One model was trained using data from the balanced group for each 3-stick combination in turn and the accuracy of each model, in distinguishing poor smell from good smell, assessed using an independent composite validation set formed by combining the Discovery iRBD, Marburg iRBD and Tracking PD datasets. The process of model training and validation was undertaken 10 times using randomly balanced training groups resulting in a total of 5600 trained models. The AUC values (computed on test dataset) associated with our 3-Sniffin'-stick model were compared individually against those associated with all other possible 3-stick combinations using pairwise t tests.

Paragraph 19

OBJECTIVE 6: To evaluate the effect of training dataset composition on 3-Sniffin'-stick model accuracy

The aforementioned models were trained using data from individuals at extremes; individuals with PD and poor smell, and controls with good smell. To compare the effects of training dataset composition, single randomly balanced groups of a) individuals with PD with poor smell/controls with good smell b) controls with poor smell/good smell, c) individuals with PD with poor smell/good smell were used to train models which were then tested according to their ability to predict poor smell in individuals with iRBD. Kolmogorov–Smirnov tests were used to compare the resultant AUC distributions.

Paragraph 20

OBJECTIVE 7: To Investigate a staged screening model to distinguish individuals with iRBD from controls

Paragraph 21

Using baseline iRBD data and longitudinal control data the sensitivity, specificity, positive predictive value (PPV) and negative predictive values (NPV) of age≥55 , RBDSQ≥5 and PD poor smell as predicted by the 3-stick model, in isolation and combination, in differentiating between individuals with iRBD and controls was assessed.

Paragraph 22

Data availability

De-identified participant data relating to the Oxford Discovery Cohort may be requested by means of a formal application to the OPDC Data Access Committee by any qualified investigator. The application form, protocol, and terms and conditions may be found at the website: opdc.medsci.ox.ac.uk/external-collaborations.

Results

Demographics

Baseline characteristics including smell status are shown in Table 1. No difference in sex was observed within disease groups. However, participants with Parkinson's in the Tracking cohort were on average older and of longer disease duration at the time of their smell assessment. Whilst participants in Tracking had lower MDS-UPDRS part 3 scores, motor assessments predated smell assessments by up to 6 months and were absent in 10%. Marburg participants were similar in age and subjective smell status to those in Discovery but had a longer disease duration at the time of their smell assessment. Self-reported poor smell concurred with poor smell on objective testing in 31% of controls, 68% of individuals with iRBD and 66% of those with PD. Within disease groups there was no significant difference in smell status on objective testing between the Discovery, Tracking and Marburg cohorts. Across Discovery groups, individuals with Parkinson's and iRBD were more likely to have poor smell than controls ($p<0.001$) but there was no difference between PD and iRBD disease groups ($p=0.67$). We found no evidence, of a difference in the total Sniffin' stick score of individuals with PD tested at 0-3 months following the start of the study and those tested at 9-12 months ($p=0.26$) or 15-18 months ($p=0.10$), to suggest a change in Sniffin' stick performance over time due to a gradual decline in odor intensity.

Paragraph 23

RESULT 1: Anise, Licorice and Banana were the 3 sticks with the greatest importance in predicting poor smell

Paragraph 24

Independent of disease group, Sniffin' sticks differed in their rates of correct identification (see Figure 2, Data available from Dryad (Table e-2)). Independent of smell status Orange,

Paragraph 24

Peppermint and Fish were most frequently (poor smell: ≥55%, good smell: ≥90%, overall: ≥70%) identified correctly. Overall, Lemon, Turpentine and Apple had the highest rates of misidentification. Within individuals of the same smell status (either poor or good sense of smell), there were significant ($p<0.01$) differences in the rates of correct Sniffin' stick identification between those with Parkinson's and controls (Data available from Dryad (Table e-2)). Anise, Licorice and Banana demonstrated the greatest differences in rates of correct identification ($p<0.001$) (Data available from Dryad (Table e-2)) and were also the top 3 MLA-identified sticks across 5340 trained models (Figure 3A).

RESULT 2: Classification accuracy improved with incremental stick number

Paragraph 25

The effect on classification accuracy of incremental cumulative Sniffin' stick number is shown in Figure 3B. The improvement in AUC from the addition of two sticks to that of a single stick, was more than three times the improvement in AUC associated with the cumulative addition of all other 13 sticks.

RESULT 3: An overall AUC≥0.90 in distinguishing poor smell from good smell using the top 3 Sniffin' sticks was replicated in a total of 3 different independent PD and iRBD cohorts (Table 2)

Paragraph 26

The AUC of 0.95 in the development dataset equated to a sensitivity, specificity, PPV and NPV of 90%, 92%, 92% and 90% respectively, assuming a probability threshold set at 0.5. A single randomly selected MLA-identified top 3 Sniffin' stick model distinguished poor smell from good smell in baseline iRBD data from individuals within the Discovery study with an

AUC of 0.90. Separately, the AUC was 0.90 in the independent Tracking (PD) cohort and 0.95 in the Marburg (iRBD) cohort (Table 2).

RESULT 4: The 3 top MLA-identified sticks (Anise, Licorice and Banana) outperformed previously described stick combinations

Paragraph 27

For the detection of poor smell, the Q stick combination (Cloves, Coffee, Rose)[27] was outperformed, when using either an absolute cut off of ≤2 correctly identified sticks or when applying trained MLAs that utilised the individual answer for each stick (out of 4 possible options) (Figure 4). Other stick combinations described by Casjens et al.,[26] Boseveldt et al.[29] and Muller et al.[25] were similarly outperformed (Figure 3B).

RESULT 5: No other 3 stick combination statistically outperformed the 3 top MLA-identified sticks

Paragraph 28

When validated against the composite validation dataset (combining Discovery iRBD, Marburg iRBD and Tracking PD datasets), the AUC associated with our top 3-stick combination was 0.90 (0.89-0.91) compared to 0.68 (0.67-0.70) for the worst predicted 3-stick combination (Turpentine, Garlic and Apple). No other 3-stick combination was statistically better than our 3 stick (Anise, Licorice and Banana) combination, in detecting poor smell ($p>0.05$). There was a significant difference ($p<0.001$) between the 3-stick combinations that included at least one of Anise/Licorice/Banana compared to those that did not include any of the 3 aforementioned sticks.

RESULT 6: Models trained using a combination of PD poor smell/control good smell were better at identifying poor smell in individuals with iRBD than those trained using PD poor smell/PD good smell or Control poor smell/Control good smell data (Data available from Dryad (Figure e-1))

Paragraph 29

The overall AUC value (for distinguishing poor smell from good smell) across 5340 trained models using the same top 3 MLA-identified sticks and data from individuals with PD and poor smell and controls with good smell was 0.95; greater than that associated with models trained using control poor smell/control good smell data (0.81) or models trained using PD poor smell/PD good smell data (0.86). When models were trained using all 560 possible 3-stick combinations, higher AUC values were again obtained when using data from individuals with PD poor smell/control good smell as opposed to control poor smell/control good smell or PD poor smell/PD good smell ($p<0.0001$).

RESULT 7: A two-step screening test comprising a) an RBDSQ≥5 and b) poor smell as predicted using the 3-Sniffin'-stick model, distinguished individuals with iRBD from controls with a sensitivity of 65%, specificity of 100%, PPV of 100% and NPV of 30% (Table 3)

Paragraph 30

Whilst the sensitivity of the two step screening test was lower than that associated with using age≥55, RBDSQ≥5 or 3-Sniffin'-stick model predicted poor smell, alone or in combination, the two-step screening test was associated with a higher specificity and PPV; its accuracy values not benefiting further from the addition of age≥55. However, the PPV is a function of the background prevalence of iRBD which is unrealistically high in our sample due to its design. If one takes the prevalence of 1.06% from a community survey in Switzerland[44] and factors in the uncertainty around our specificity (95% lower confidence

interval 90% and substituting 99% for the upper confidence interval of 100%, as a specificity of 100% will always yield a PPV of 100%) then the PPV could range from 7% to 41% if the specificity ranges from 90 to 99%, compared to a PPV range of 3% to 12% when using the RBDSQ alone.

Discussion

Paragraph 31

We describe the application of a 3-Sniffin'-stick test to a total of 1933 individuals with iRBD, PD and controls; to our knowledge, the largest study of its kind. Poor sense of smell on objective testing was evident in 80% of individuals with Parkinson's, 72% of individuals with iRBD and 15% of controls, surpassing rates of self-reported poor sense of smell of 61%, 58% and 11% respectively. The three sticks with the highest (Orange, Peppermint and Fish)[18] and lowest (Lemon, Turpentine, Apple)[24] rates of correct identification across groups matched those reported previously.

Paragraph 32

Our aim was to develop an abbreviated Sniffin' stick test for the detection of impaired olfaction in prodromal and manifest PD, with the assumption that individuals with iRBD, prodromal to PD, lie on a continuum between controls and those with PD. We did not seek to distinguish individuals with PD from controls on the basis of their sense of smell alone; rather to identify those who had a poor sense of smell based on normative data specific to age and sex. Models were therefore trained with smell data at either extreme; data from controls with good smell (normosmia and super smellers) and data from individuals with PD and poor smell (hyposmia and functional anosmia). When tested on independent data from individuals with iRBD, models trained using PD poor smell/control good smell data were better at identifying poor smell in individuals with iRBD than those trained using control good smell/control poor smell data or PD good smell/PD poor smell data, suggesting a pattern of smell loss that is also recapitulated in individuals with iRBD who have prodromal parkinsonism (Data available from Dryad (Figure e-1)).

Paragraph 33

These findings are consistent with previous studies in which we and others have shown that individuals with iRBD have a non-motor profile comparable to those with PD, with similarities in cognition, depression, anxiety, apathy, impulsive compulsive behaviors, sleep and autonomic dysfunction.[2, 12, 34, 45-48] Furthermore, individuals with iRBD show subtle motor dysfunction, compared to age and gender matched controls, which is less than that required to meet the diagnosis of PD.[34] This emerging wealth of phenotypic data suggests that the vast majority of individuals with iRBD already manifest motor and non-motor features of prodromal parkinsonism, sitting on one end of a disease spectrum, on a continuum with established PD.

Paragraph 34

The cut-off points used to categorise olfactory performance were based on normative data stratified by age and gender and so our models were not adjusted for either of the aforementioned factors. Instead we utilised the raw answers provided for each stick by each participant, which may in part explain the less than perfect 16 stick AUC of 0.99 calculated using validation data excluded from the training process. Nonetheless, despite their exclusion as model input variables, the subgroup performance of the top 3-stick trained models by age and sex was largely equivalent (Data available from Dryad (Table e-3)); the exception being for the 31 to 40 age group where the associated dataset was small.

Paragraph 35

When sticks were ranked in descending order, the sharp reduction in predictor importance seen after the top 3 sticks (Anise, Licorice and Banana), translated into a relative reduction in AUC improvement with models utilising 4 or more sticks (Figure 3). Cumulative incorporation of Sniffin' sticks into models (Figure 3B), according to their data-driven

ranking resulted in AUCs that outperformed previously reported 3 and 5 Sniffin' stick combinations.[25-27, 29] The Casjens et al. and Boesveldt et al. 3-stick combinations yielded AUCs that were slightly lower than those associated with our top 3 MLA-identified sticks; perhaps unsurprising given that Anise (the stick with the highest MLA-identified predictor importance) was common to all three top 3 combinations and the Boesvelt et al. combination additionally had Licorice in common (Data available from Dryad (Table e-1)).

Paragraph 35

Indeed, the S8 (8-stick) combination described by Mahlknecht et al. which interestingly had 5 out of 8 sticks in common including the top 3 sticks, was associated with an AUC virtually identical to that of the top 8 MLA-identified sticks.[24] Furthermore, comparing the top 3 MLA-identified sticks to all other 559 possible 3-stick combinations (including those previously described) through the application of trained models to the independent composite validation dataset, no other 3-stick combination was statistically better than our 3-stick combination in detecting poor smell ($p>0.05$).

Paragraph 36

Having derived a top 3 MLA-identified stick test of poor smell, we assessed its value both alone and in combination with age≥55 and RBDSQ≥5 in screening for individuals with iRBD. In keeping with the main body of scientific literature, as a single test, the RBDSQ was associated with excellent sensitivity.[49] However, the combination of RBDSQ≥5 and poor smell identified from 3-stick testing, yielded 100% specificity, though the lower 95% confidence interval was equal to 90%. Depending on the true specificity and community prevalence of iRBD, the PPV could range widely from 7% to 41% (compared to 3% to 12% when using the RBDSQ alone) but these figures, if replicated, suggest a potential two-pronged screening test for community cases of subclinical iRBD that could be used to facilitate large-scale screening of individuals as part of clinical trials for prodromal PD. The

addition of age≥55 did not result in any further improvement in accuracy values. Individuals with polysomnographically confirmed iRBD who did not fulfil both criteria (RBDSQ≥5 and poor smell on 3-Sniffin'-stick test) were younger (mean (SD) age 61.7 (10.4) versus 66.3 (7.5) p<0.001) and had a lower MDS-UPDRS III score (mean (SD) 5.5 (5.3) versus 3.8 (3.2) p=0.01) compared to those who met both criteria. There was no difference in sex (p=0.10) between the two subgroups. Though not created to detect disease *per se*, the 3-stick model trained to detect individuals with poor smell, when used in isolation, distinguished iRBD from controls with a sensitivity of 67% and a specificity of 80%; comparing favourably to respective values of 56% and 89% reported by Huang et al. wherein the abbreviated 5 Sniffin' stick test was created with the specific intent of distinguishing disease groups.[23]

Paragraph 36

One of the main weaknesses of this study and in the evaluation of the 3-stick test in the screening of iRBD, was the absence of an independent control cohort. Accuracy values were therefore calculated using longitudinal control data that was treated as independent of the baseline control data. Additionally, in keeping with iRBD studies worldwide, individuals with iRBD who had been recruited into the Discovery and Marburg studies were those who had originally presented to their clinician for evaluation of their sleep disturbance; as such they may be expected to present with a more severe phenotype compared to individuals detected on population screening. Answers provided on the RBDSQ may also have been affected by foreknowledge of their iRBD diagnosis. Future work will apply our methods to derive a population of individuals with community ascertained iRBD through the application of the 3-stick/RBDSQ test, acknowledging that the community prevalence of iRBD will affect the associated PPV and understanding the potential for regional variations to the largely German-derived normative data from which olfactory performance was categorised.

Paragraph 37

Paragraph 37

Although the findings of our study are consistent across three large independent cohorts, future work will also explore its application to other international cohorts where cultural and regional variations in exposure to Anise and Licorice, both similar in smell, may affect the generalisability of our results beyond the relatively homogeneous European cohorts described. A further interesting avenue of exploration includes the evaluation of longitudinal changes in the 3-stick test and its relation to clinical outcomes of interest.

Paragraph 38

In conclusion, we demonstrate that a 3-stick model comprising the Sniffin' sticks Anise, Licorice and Banana detects olfactory dysfunction with high levels of accuracy in individuals with Parkinson's and iRBD. Its ease of administration and relative cost effectiveness supports a role in screening for iRBD and in clinical phenotyping, where prognostication may be facilitated through standardised assessment of olfactory dysfunction.

Acknowledgements

We are grateful to participants in the Discovery, Tracking and Marburg studies, without whom, none of this work would have been possible. Additionally, we would like to thank members of the research teams at the various sites for their faithful work in clinical phenotyping.

Tables

Table 1 | Baseline demographics of participants in the Discovery study alongside the independent Tracking and Marburg datasets.

	Control	PD			RBD		
	Discovery	Discovery	Tracking	P[a]	Discovery	Marburg	P[a]
Baseline demographics							
N	313	890	452	-	241	37	-
Age	64.4 (9.8)	66.5 (9.6)	68.0 (9.0)	<0.01	64.6 (8.9)	67.3 (8.8)	0.97
Male sex	165 (53%)	569 (64%)	290 (64%)	0.93	211 (88%)	34 (87%)	0.78
Disease duration at Sniffin' stick assessment	-	1.2 (0.9)	1.9 (1.0)	<0.001	1.4 (1.8)	3.5 (2.9)	<0.001
MDS-UPDRS part 3	1.8 (2.5)[b]	26.4 (10.8)	23.4 (12.8)[b]	<0.001	5.1 (5.9)[c]	5.0 (2.2)[d]	0.95
Montreal Cognitive Assessment	26.7 (2.5)[e]	25.0 (3.3)[e]	25.4 (3.4)[e]	0.03	25.1 (2.9)[e]	27.5 (2.1)[e]	<0.001
Subjective poor smell	34 (11%)[f]	540 (61%)	-	-	139 (58%)[f]	20 (59%)[g]	0.66
Objective smell testing							
Mean total Sniffin' score (SD)	12.1 (2.3)	7.1 (2.9)	7.4 (3.0)	0.04	7.9 (3.2)	7.5 (2.7)	0.44

Poor smell	Functional anosmia	25 (8%)	623 (70%)	309 (68%)	0.16	147 (61%)	24 (65%)	0.35
	Hyposmia	21 (7%)	98 (11%)	48 (11%)		23 (10%)	6 (16%)	
Good smell	Normosmia	217 (69%)	161 (18%)	84 (19%)		64 (27%)	7 (19%)	
	Super smell	50 (16%)	8 (1%)	11 (2%)		7 (3%)	0	

Table 1 legend (missing data):

[a]p value determined using a two-sample t-test or chi squared test. [b]2 Control participants, 7 RBD participants and 11 PD participants from the Discovery study had an incomplete MDS-UPDRS part 3 score. Descriptive statistics relating to the MDS-UPDRS part 3 score for the aforementioned groups were therefore calculated across those with complete scores. [c]In the Tracking cohort MDS-UPDRS part 3 scores were assessed up to 6 months before the Sniffin' sticks test was performed. 45 participants did not have a motor examination within the 6 month window. [d]Motor impairment was assessed in the Marburg cohort using the original version of the UPDRS (as opposed to the MDS-UPDRS used in Discovery and Tracking). Scores were converted using the formula: (original UPDRS part 3 score x 1.2) + 2.3 which was developed to convert scores from individuals with PD at Hoehn & Yahr stage I&II.[50] Additionally 1 of the Marburg participants had their motor assessment 7 months after their smell assessment (as opposed to within a week of their assessment). [e]A total MoCA score was not available for 13 Controls, 37 participants with PD and 2 participants RBD in Discovery, 41 participants with PD in the Tracking cohort and 1 participant in the Marburg cohort. Summary statistics are calculated across those with available scores. Subjective

smell status was missing in 1 participant in the Discovery control and RBD groups[f] and 2 participants in the Marburg RBD group.[g] It was not assessed in the Tracking study.

Table 2 | Comparison of area under the curve values for MLA-trained 3-Sniffin'-stick models in the detection of hyposmia or functional anosmia.

	AUC (95% confidence interval)		
	16 sticks	MLA-identified 3 sticks	3 Q sticks
Development			
Overall PD poor smell model accuracy[a]	0.99 (0.99-0.99)	0.95 (0.95-0.96)	0.87 (0.86-0.88)
Validation			
Internal: RBD baseline[b]	-	0.90 (0.85-0.94)	0.83 (0.76-0.88)
External: PD Tracking[b]	-	0.90 (0.85-0.93)	0.81 (0.76-0.86)
External: RBD Marburg[b]	-	0.95 (0.82-0.99)	0.82 (0.58-0.97)

[a] As calculated across 5340 trained models, following the application of a leave one subject out cross validation scheme, where the data used to train models and the data used to assess model accuracy were mutually exclusive. [b]As predicted by a single randomly selected model trained using baseline PD/Control Discovery data for each 3-stick combination (MLA-identified or previously described Q stick) with the same model being applied to both internal and external validation datasets. Confidence intervals for the validation datasets are calculated across single predictions for each set of clinical data whereas confidence intervals for the development dataset are calculated using a bootstrapping approach across

all 5340 predictions from the trained models from the 10 balanced datasets; the two confidence intervals are therefore not directly comparable.

Table 3 | RBD/Control detection accuracies associated with age≥55[a], RBDSQ≥5[b] and MLA-identified 3-Sniffin'-stick predicted hyposmia/functional anosmia[c], alone and in combination.

	Sensitivity	Specificity	PPV	NPV
Age[a]	88%	14%	87%	16%
RBDSQ[b]	97%	80%	97%	82%
Sniffin'[c]	67%	80%	96%	27%
Age[a] + Sniffin'[c]	63%	83%	96%	26%
Age[a] + RBDSQ[b]	86%	89%	98%	49%
RBDSQ[b] + Sniffin'[c]	65%	100%	100%	30%
RBDSQ[b] + Age[a] + Sniffin'[c]	62%	100%	100%	28%

Figure legends (figures uploaded separately)

Figure 1 | Flow charts demonstrating the data used to train poor smell/good smell models

Figure 2 | Spider web plot demonstrating the proportion (%) of individuals in each group who correctly identified each Sniffin' stick.

Each Sniffin' stick is represented by a line radiating out from the centre of the plot with points at the maximal radius indicating 100% correct identification

Figure 3 | A) The average predictor importance of each Sniffin' stick across 5340 PD poor smell/control good smell models. B) The AUC associated with incremental Sniffin' stick combinations compared to those previously described.

Figure 3A: Predictor importance is derived from the random forest algorithm using Gini's diversity index criteria for binary splitting. Figure 3B legend: Blue line with unfilled dots: Discovery trained PD poor smell/control good smell models; Pink filled in dot: Q stick Hummel et al. (3-stick) combination; Green asterix: Casjens et al. (3-stick) combination; Orange square: Boesveldt et al. (3-stick) combination; Purple pentagon: Mueller et al. BSIT (5-stick) combination; Magenta cross: Mahlknecht et al. S8 (8-stick) combination.

Figure 4 | Petal plots demonstrating the sensitivity, specificity, positive predictive value and negative predictive values associated with using an absolute score of ≤2 out of 3 sticks (A,B,E,F) or a trained MLA that takes into account each individual option for

each stick (C,D,G,H) with sticks being chosen empirically (A,C,E,G) or by MLA predictor importance ranking (B,D,F,H).

The larger the petal size, the greater the accuracy, with the tip of each petal indicating the percentage accuracy for each summary measure. Accuracy values displayed for the trained MLA relate to a cut off probability threshold of 0.5.

References

1. Doty RL. Olfactory dysfunction in Parkinson disease. Nature reviews Neurology 2012;8:329-339.

2. Baig F, Lawton M, Rolinski M, et al. Delineating nonmotor symptoms in early Parkinson's disease and first-degree relatives. Movement disorders : official journal of the Movement Disorder Society 2015;30:1759-1766.

3. Morley JF, Cheng G, Dubroff JG, Wood S, Wilkinson JR, Duda JE. Olfactory Impairment Predicts Underlying Dopaminergic Deficit in Presumed Drug-Induced Parkinsonism. Mov Disord Clin Pract 2017;4:603-606.

4. Yang HJ, Kim YE, Yun JY, Ehm G, Kim HJ, Jeon BS. Comparison of sleep and other non-motor symptoms between SWEDDs patients and de novo Parkinson's disease patients. Parkinsonism Relat Disord 2014;20:1419-1422.

5. Domellof ME, Lundin KF, Edstrom M, Forsgren L. Olfactory dysfunction and dementia in newly diagnosed patients with Parkinson's disease. Parkinsonism Relat Disord 2017;38:41-47.

6. Fullard ME, Tran B, Xie SX, et al. Olfactory impairment predicts cognitive decline in early Parkinson's disease. Parkinsonism Relat Disord 2016;25:45-51.

7. Schrag A, Siddiqui UF, Anastasiou Z, Weintraub D, Schott JM. Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. The Lancet Neurology 2017;16:66-75.

8. Ham JH, Lee JJ, Sunwoo MK, Hong JY, Sohn YH, Lee PH. Effect of olfactory impairment and white matter hyperintensities on cognition in Parkinson's disease. Parkinsonism Relat Disord 2016;24:95-99.

9. Gjerde KV, Muller B, Skeie GO, Assmus J, Alves G, Tysnes OB. Hyposmia in a simple smell test is associated with accelerated cognitive decline in early Parkinson's disease. Acta Neurol Scand 2018;138:508-514.

10. Kang SH, Lee HM, Seo WK, Kim JH, Koh SB. The combined effect of REM sleep behavior disorder and hyposmia on cognition and motor phenotype in Parkinson's disease. Journal of the neurological sciences 2016;368:374-378.

11. Heinzel S, Berg D, Gasser T, Chen H, Yao C, Postuma RB. Update of the MDS research criteria for prodromal Parkinson's disease. Movement disorders : official journal of the Movement Disorder Society 2019.

12. Postuma RB, Iranzo A, Hu M, et al. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. Brain : a journal of neurology 2019.

13. Postuma RB, Gagnon JF, Bertrand JA, Genier Marchand D, Montplaisir JY. Parkinson risk in idiopathic REM sleep behavior disorder: preparing for neuroprotective trials. Neurology 2015;84:1104-1113.

14. Shill HA, Hentz JG, Caviness JN, et al. Unawareness of Hyposmia in Elderly People With and Without Parkinson's Disease. Mov Disord Clin Pract 2016;3:43-47.

15. Hummel T, Sekinger B, Wolf SR, Pauli E, Kobal G. 'Sniffin' sticks': olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. Chem Senses 1997;22:39-52.

16. Doty RL, Shaman P, Kimmelman CP, Dann MS. University of Pennsylvania Smell Identification Test: a rapid quantitative olfactory function test for the clinic. Laryngoscope 1984;94:176-178.

17. Haehner A, Mayer AM, Landis BN, et al. High test-retest reliability of the extended version of the "Sniffin' Sticks" test. Chem Senses 2009;34:705-711.

18. Hummel T, Konnerth CG, Rosenheim K, Kobal G. Screening of olfactory function with a four-minute odor identification test: reliability, normative data, and investigations in patients with olfactory loss. Ann Otol Rhinol Laryngol 2001;110:976-981.

19. Lawton M, Hu MT, Baig F, et al. Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease. Parkinsonism Relat Disord 2016;33:96-101.

20. Boesveldt S, Verbaan D, Knol DL, van Hilten JJ, Berendse HW. Odour identification and discrimination in Dutch adults over 45 years. Rhinology 2008;46:131-136.

21. Morley JF, Cohen A, Silveira-Moriyama L, et al. Optimizing olfactory testing for the diagnosis of Parkinson's disease: item analysis of the university of Pennsylvania smell identification test. NPJ Parkinson's disease 2018;4:2.

22. Crespo Cuevas AM, Ispierto L, Vilas D, et al. Distinctive Olfactory Pattern in Parkinson's Disease and Non-Neurodegenerative Causes of Hyposmia. Neurodegener Dis 2018;18:143-149.

23. Huang SF, Chen K, Wu JJ, et al. Odor Identification Test in Idiopathic REM-Behavior Disorder and Parkinson's Disease in China. PloS one 2016;11:e0160199.

24. 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.

25. 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.

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

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

28. Campabadal A, Segura B, Junque C, et al. Comparing the accuracy and neuroanatomical correlates of the UPSIT-40 and the Sniffin' Sticks test in REM sleep behavior disorder. Parkinsonism Relat Disord 2019;65:197-202.

29. 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.

30. Miyamoto T, Miyamoto M, Iwanami M, Hirata K. Olfactory dysfunction in Japanese patients with idiopathic REM sleep behavior disorder: comparison of data using the university of Pennsylvania smell identification test and odor stick identification test for Japanese. Movement disorders : official journal of the Movement Disorder Society 2010;25:1524-1526.

31. Miyamoto T, Miyamoto M, Iwanami M, et al. Olfactory dysfunction in idiopathic REM sleep behavior disorder. Sleep medicine 2010;11:458-461.

32. Krismer F, Pinter B, Mueller C, et al. Sniffing the diagnosis: Olfactory testing in neurodegenerative parkinsonism. Parkinsonism Relat Disord 2017;35:36-41.

33. Stiasny-Kolster K, Mayer G, Schafer S, Moller JC, Heinzel-Gutenbrunner M, Oertel WH. The REM sleep behavior disorder screening questionnaire--a new diagnostic instrument. Movement disorders : official journal of the Movement Disorder Society 2007;22:2386-2393.

34. Barber TR, Lawton M, Rolinski M, et al. Prodromal Parkinsonism and neurodegenerative risk stratification in REM sleep behaviour disorder. Sleep 2017.

35. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of neurology, neurosurgery, and psychiatry 1992;55:181-184.

36. Lawton M, Ben-Shlomo Y, May MT, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. Journal of neurology, neurosurgery, and psychiatry 2018;89:1279-1287.

37. Swallow DM, Lawton MA, Grosset KA, et al. Statins are underused in recent-onset Parkinson's disease with increased vascular risk: findings from the UK Tracking Parkinson's and Oxford Parkinson's Disease Centre (OPDC) discovery cohorts. Journal of neurology, neurosurgery, and psychiatry 2016;87:1183-1190.

38. International Classification of Sleep Disorders. American Academy of Sleep Medicine 2014.

39. Oleszkiewicz A, Schriever VA, Croy I, Hahner A, Hummel T. Updated Sniffin' Sticks normative data based on an extended sample of 9139 subjects. Eur Arch Otorhinolaryngol 2019;276:719-728.

40. Kobal G, Klimek L, Wolfensberger M, et al. Multicenter investigation of 1,036 subjects using a standardized method for the assessment of olfactory function combining tests of odor identification, odor discrimination, and olfactory thresholds. Eur Arch Otorhinolaryngol 2000;257:205-211.

41. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor Newsl 2004;6:1–6.

42. Drummond C, Holte R. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats OverSampling. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets 2003.

43. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

44. Haba-Rubio J, Frauscher B, Marques-Vidal P, et al. Prevalence and Determinants of REM Sleep Behavior Disorder in the General Population. Sleep 2017.

45. Baig F, Kelly MJ, Lawton MA, et al. Impulse control disorders in Parkinson disease and RBD: A longitudinal study of severity. Neurology 2019;93:e675-e687.

46. Baig F, Lawton MA, Rolinski M, et al. Personality and addictive behaviours in early Parkinson's disease and REM sleep behaviour disorder. Parkinsonism Relat Disord 2017;37:72-78.

47. Rolinski M, Zokaei N, Baig F, et al. Visual short-term memory deficits in REM sleep behaviour disorder mirror those in Parkinson's disease. Brain : a journal of neurology 2016;139:47-53.

48. Szewczyk-Krolikowski K, Tomlinson P, Nithi K, et al. The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. Parkinsonism Relat Disord 2014;20:99-105.

49. Li K, Li SH, Su W, Chen HB. Diagnostic accuracy of REM sleep behaviour disorder screening questionnaire: a meta-analysis. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology 2017;38:1039-1046.

50. Goetz CG, Stebbins GT, Tilley BC. Calibration of unified Parkinson's disease rating scale scores to Movement Disorder Society-unified Parkinson's disease rating scale scores. Movement Disorders 2012;27:1239-1242.