
Supplementary information

Remote smartphone monitoring of Parkinson's disease and individual response to therapy

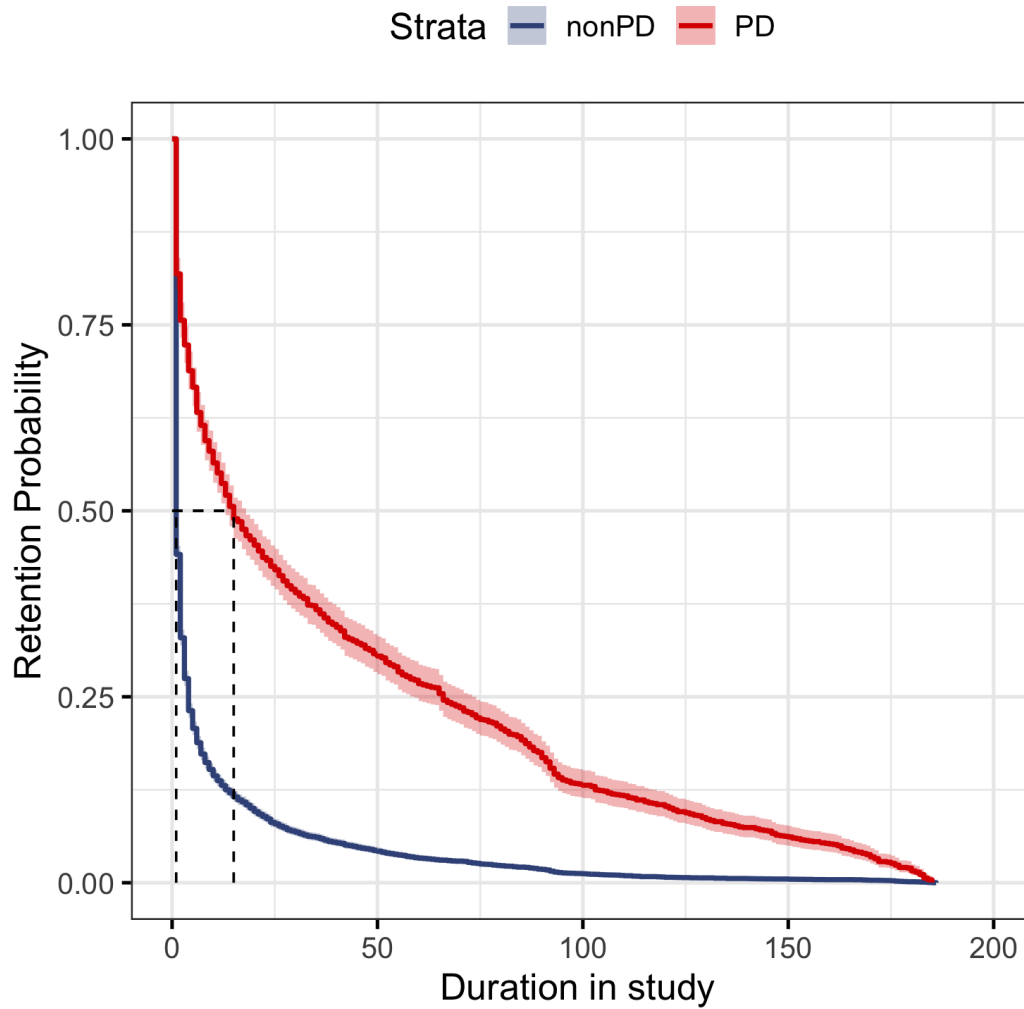
In the format provided by the
authors and unedited

Supplement

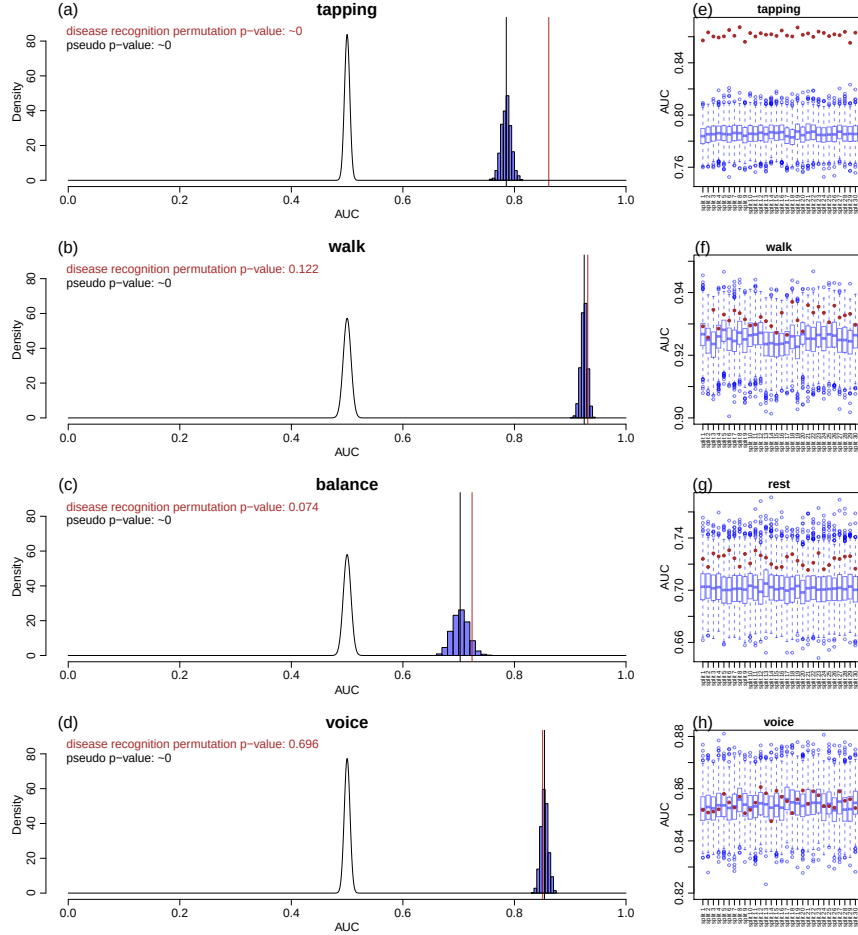
Contents

1	Supplementary Figures 1 to 10	2
2	Supplementary Note 1	12
2.1	Supplementary Figures 11 to 30	17
3	Supplementary Figure 31	37

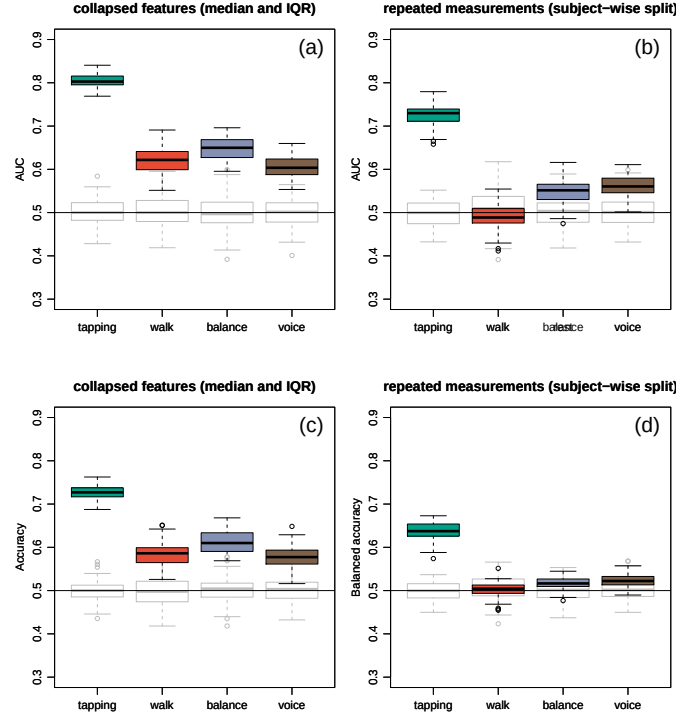
1 Supplementary Figures 1 to 10



Supplementary Figure 1: Participant retention in the mPower study. The figure shows the retention curves for the PD and non-PD participants, based on the total duration of time in the study as the outcome metric. The median total duration in the study was of 15 days for PD participants, and 1 day for non-PD participants.

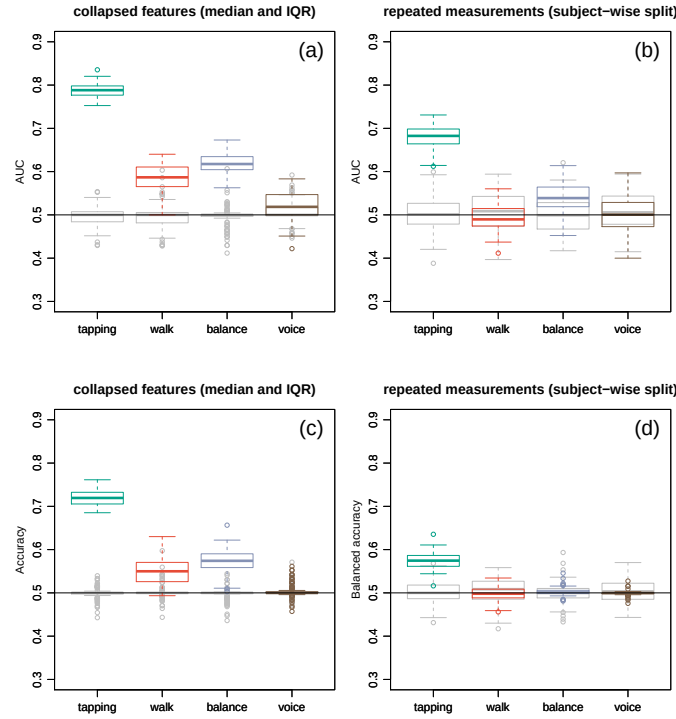


Supplementary Figure 2: Identity confounding arising from record-wise splits of the mPower data. Panels a-d illustrate the presence of a high degree of identity confounding in the tapping, walk, rest, and voice tasks, respectively. The blue histograms present the disease recognition null distributions (Online Methods), generated by computing the AUC statistic in shuffled versions of the data where the disease labels were permuted in a subject-wise manner (i.e., all the records of a given participant are either assigned the PD or the non-PD label), using data split in a record-wise manner (where, for each participant, a random subset of the records contributed by the participant are assigned to the training set, while the remaining records are assigned to the test set). The fact that the null distributions are centered away from 0.5 (the random guess baseline value for the AUC metric) indicate the presence of identity confounding. (Note that the subject-wise label shuffling strategy breaks the association between labels and feature data, but still preserves the association between features and participant identities, so that it neutralizes the classifier ability to recognize disease, but still allows the classifier to identify subjects). The brown vertical lines represent the AUC value computed in the original (un-shuffled) data. Except for tapping, the observed AUC values were located inside the range of the null distributions, suggesting that the algorithms are mostly performing subject identification (disease recognition permutation p-values equal to 0.122, 0.074, and 0.696 for the walk, rest and voice tasks, respectively). The black vertical lines represent the median of the disease recognition null distributions. The pseudo p-values (Online Methods) correspond to the area under the black density to the right of the black line, and further confirm a high degree of identity confounding in all tasks (pseudo p-value $< 1 \times 10^{-16}$ for all tasks). Panels e-h show the disease recognition null distributions (blue boxplots) and the respective observed AUC values (brown dots) for 30 distinct training/test data splits, and show that the results in panels a-d are robust with respect to the random data split.



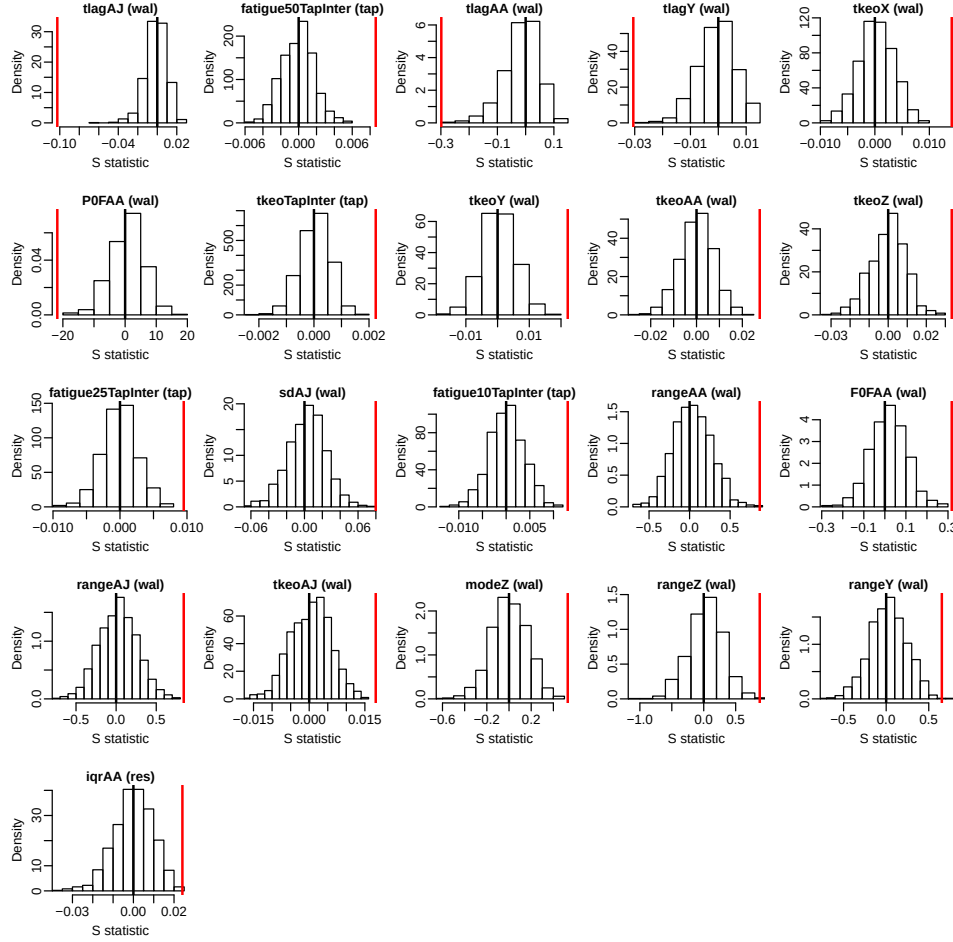
Supplementary Figure 3: PD vs non-PD classification results for random forest.

Panel a reports the AUC distributions from the PD vs non-PD classification for the tapping, walk, balance, and voice activities, based on collapsed features (i.e., where, for each extracted feature, all the record measurements contributed by a participant were summarized into the median and the interquartile range of the repeated measurements). The boxplots represent the AUC values computed from 100 random splits of the data into training and test sets, with half of the data used for training and half for testing (the random splits were stratified by label). (For each boxplot, the box corresponds to the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.) Panel b reports the corresponding AUC distributions from classifiers built with repeated measurements (multiple records) contributed by each participant, but where the training and test sets were generated using subject-wise random splits of the data (in order to avoid the identity confounding artifact). The boxplots were based on 100 subject-wise random splits of the data. Except for tapping, the classification performance was modest for all tasks. We adopted balanced accuracy for the repeated measurement analyses (panel d) due to unbalance in the PD/non-PD classes (since participants contributed varying numbers of records). All the analyses were based in the subset of matched participants that contributed at least 5 records as described in the Online Methods. For panels a and c the sample sizes corresponds to the number of participants, and were equal to 960, 464, 466, and 620, for the tapping, walk, balance, and voice activities, respectively. (For panels b and d the sample sizes correspond to the total number of records across the participants and are given by 38,494, 36,990, 18,029, and 19,077, respectively.) In all panels the grey boxplots report the results obtained by training the classifiers on shuffled labels.

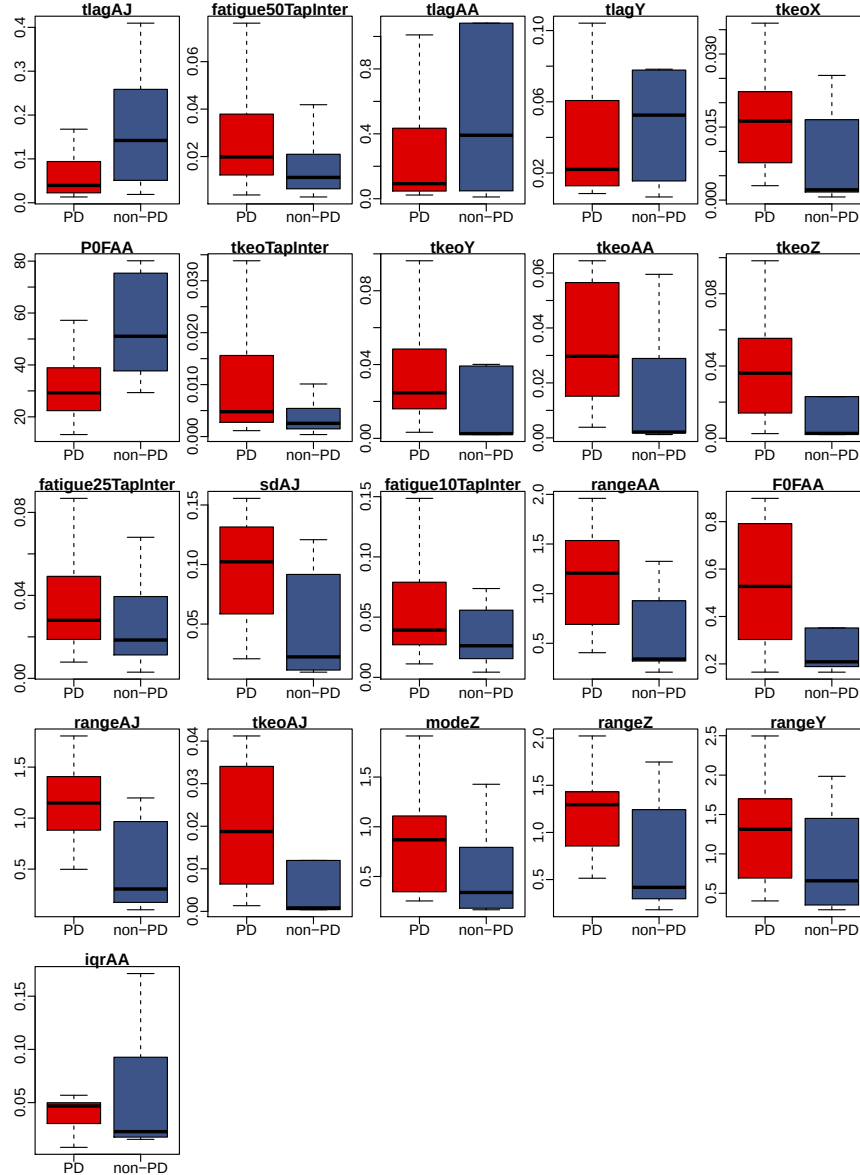


Supplementary Figure 4: PD vs non-PD classification results for ridge-regression.

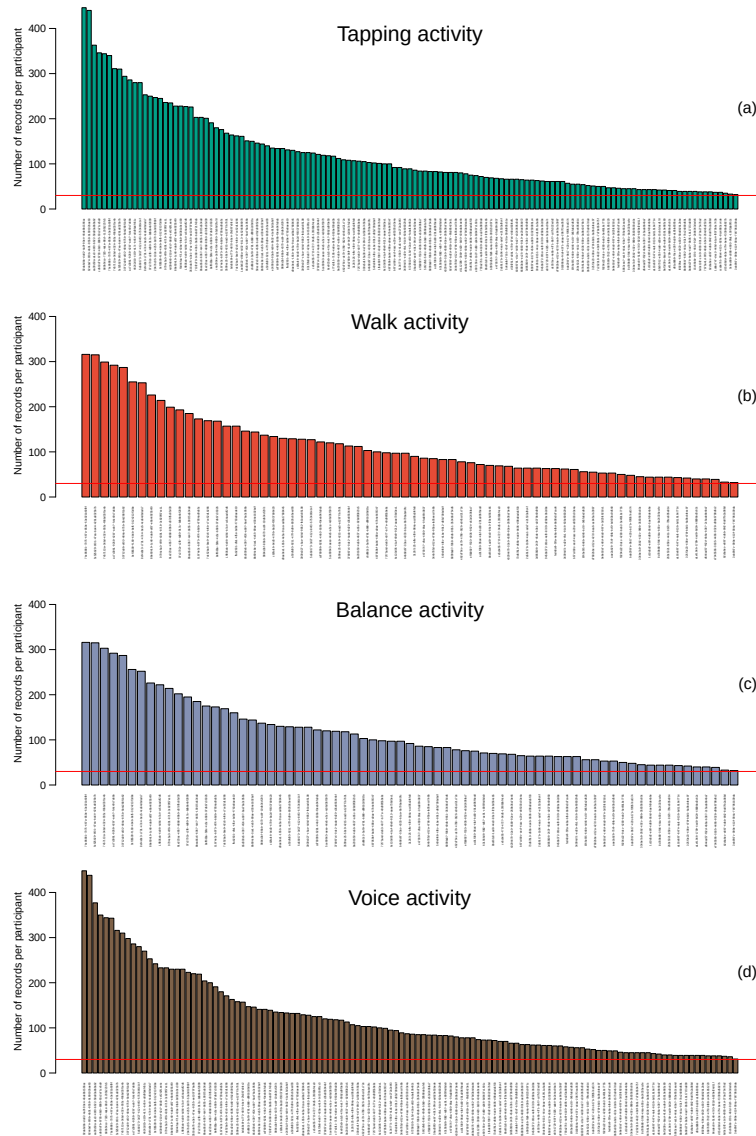
Panel a reports the AUC distributions from the PD vs non-PD classification for the tapping, walk, balance, and voice activities, based on collapsed features (i.e., where, for each extracted feature, all the record measurements contributed by a participant were summarized into the median and the interquartile range of the repeated measurements). The boxplots represent the AUC values computed from 100 random splits of the data into training and test sets, with half of the data used for training and half for testing (the random splits were stratified by label). (For each boxplot, the box corresponds to the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.) Panel b reports the corresponding AUC distributions from classifiers built with repeated measurements (multiple records) contributed by each participant, but where the training and test sets were generated using subject-wise random splits of the data (in order to avoid the identity confounding artifact). The boxplots were based on 100 subject-wise random splits of the data. Except for tapping, the classification performance was modest for all tasks. Panels c and d show the respective plots for the accuracy and balance accuracy metrics, respectively. We adopted balanced accuracy for the repeated measurement analyses (panel d) due to unbalance in the PD/non-PD classes (since participants contributed varying numbers of records). All the analyses were based in the subset of matched participants that contributed at least 5 records as described in the Online Methods. For panels a and c the sample sizes corresponds to the number of participants, and were equal to 960, 464, 466, and 620, for the tapping, walk, balance, and voice activities, respectively. (For panels b and d the sample sizes correspond to the total number of records across the participants and are given by 38,494, 36,990, 18,029, and 19,077, respectively.) In all panels the grey boxplots report the results obtained by training the classifiers on shuffled labels.



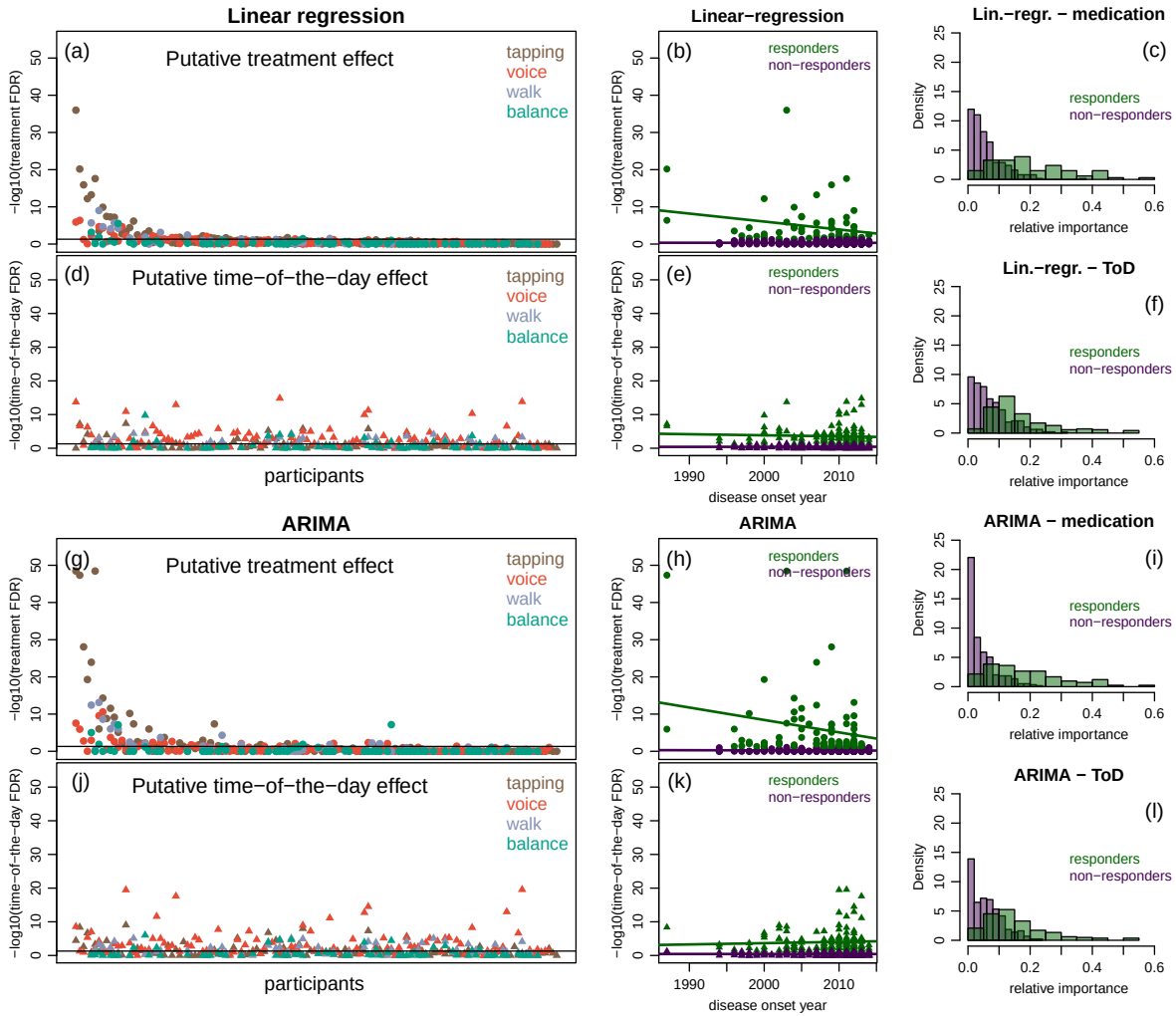
Supplementary Figure 5: Variability comparison between Parkinson patients (PD) and control (non-PD) participants. The panels show the permutation null distributions for the 21 features (out of 186) that showed a statistically significant difference between the interquartile range (IQR) distributions for PD cases and non-PD participants (after multiple testing adjustment with Benjamini-Hochberg correction). The red lines represent the observed value of the S_k statistic for feature k , $S_k = \text{median}_i\{IQR_{ik}\} - \text{median}_j\{IQR_{jk}\}$, where i and j index PD and non-PD participants, respectively. The black lines are set at 0. The title of each panel shows the feature name and the activity task. For 17 (out of 21) features we observed larger IQRs in PD participants than in non-PD participants. Results were based on 1,000 permutations. A binomial test computing the tail probability of observing 17 or more successes out of 21 trials, under the null hypothesis that the probability of success was 0.5 shows that this result is statistically significant ($p < 0.0036$). The number of PD/non-PD participants available for these analysis in the tapping, walk, balance, and voice assessments corresponded, respectively, to 67/28, 16/6, 12/5, and 65/23 participants. Feature definitions are provided in <https://github.com/Sage-Bionetworks/mpowertools/blob/master/FeatureDefinitions.md>



Supplementary Figure 6: Variability comparison between Parkinson patients (PD) and control (non-PD) participants. The panels show the IQR distributions for PD cases and non-PD participants for the 21 features that showed statistically significant differences, after multiple testing correction. The number of PD/non-PD participants available for these analysis in the tapping, walk, balance, and voice assessments corresponded, respectively, to 67/28, 16/6, 12/5, and 65/23 participants. The distributions are represented by boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

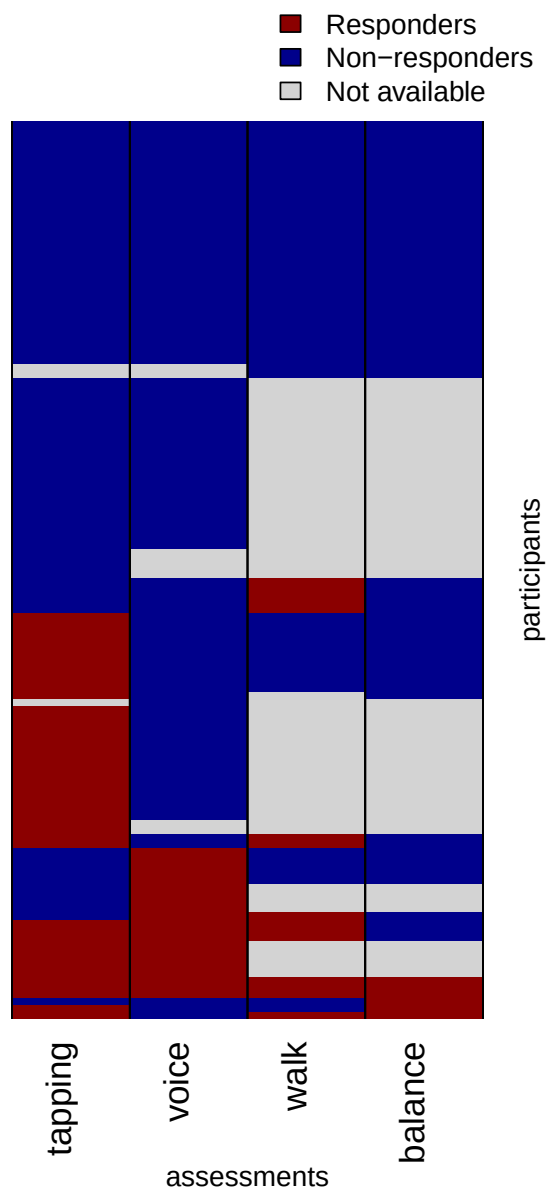


Supplementary Figure 7: Number of records contributed by each participant for each of the activity tasks. Panel a shows the counts for the tapping activity, where the number of records per participant ranged from 32 to 446, with 1st-quartile, median, mean, and 3rd-quartile given, respectively, by 55, 89, 122.1, and 151. Panel b shows the counts for the walk activity, where the number of records per participant ranged from 32 to 316, with 1st-quartile, median, mean, and 3rd-quartile given, respectively, by 61.25, 93.50, 115.36, and 145.50. Panel c shows the counts for the balance activity, where the number of records per participant ranged from 32 to 316, with 1st-quartile, median, mean, and 3rd-quartile given, respectively, by 63, 92, 116.2, and 145. Panel d shows the counts for the voice activity, where the number of records per participant ranged from 31 to 448, with 1st-quartile, median, mean, and 3rd-quartile given, respectively, by 56.75, 86.50, 123.46, and 149.50. The horizontal red line is set at 30, and shows the minimum number of records required to include a participant in the analyses.

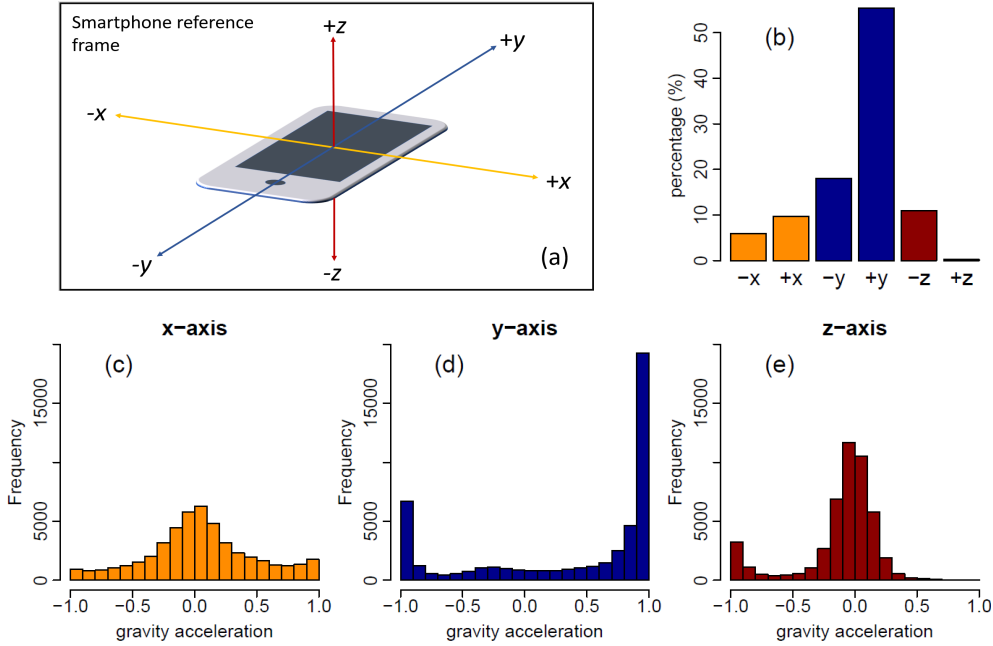


Supplementary Figure 8: Personalized fluctuations in performance across task and association with on-off motor fluctuations and time-of-the-day variations, for the standard linear regression, and the regression with ARIMA errors approaches. Panels a-f show the results for linear regression, while panels g-l report the results for the ARIMA approach. Comparison with the results presented in Figure 2 in the main text, shows that the results across the 3 approaches are largely consistent, although the Newey-West (presented in the main text) and ARIMA approaches tended to produce a larger number of statistically significant results than the standard linear regression.

Treatment responders



Supplementary Figure 9: Personalized changes in performance based on medication on-off fluctuations. Participants demonstrated on-off fluctuations differently across assessment (red indicates statistically significant putative response to L-Dopa medication, blue no-fluctuation and gray no available data) with some showing response in subset of active assessments others across multiple assessments. Fluctuations in performance was even personal within assessments where different participants showed different response rates for specific features.

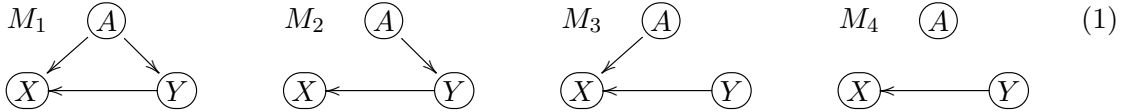


Supplementary Figure 10: Checking for instruction adherence in the walking assessment. For the walking assessment, participants were instructed to place their phones in their pockets and walk in a straight line for 30 seconds. Here, we evaluate if participants followed this instruction using a simple heuristic. The basic idea is to check the position of the phone during the activity by inspecting which coordinate in the reference frame of the phone (panel a) shows the strongest gravity acceleration. The reference frame of the phone is composed of the x, y, and z axes with positive and negative ranges as illustrated in panel a. (For instance, if the phone is left on a table with the screen facing up, the sensors will record a gravity acceleration of approximately 1g in the z-axis, and approximately 0g in the other axes. Alternatively, if the screen is facing down the table, then the sensors will record a -1g acceleration in the z-axis, and 0g in the other axes.) Now, if a participant has followed the instructions and has put the phone in the pocket we would expect to see stronger gravity accelerations in the y-axis than in the other axes (with the y-axis accelerations closer to 1g when the microphone is facing down, and closer to -1g when the phone is upside-down with the microphone facing up). Panels c, d, and e, show the distributions of the (median) gravity accelerations on the x, y, and z axes, across 47,828 walk recordings. (Note that each record captures 3 time series of gravity accelerations, one for each of the x, y, and z axes and, here, we report the distributions of the median value of each of these 3 time series of acceleration values). Panel b shows the percentages of the records whose gravity acceleration signal was closer to either 1g or -1g in each of the axes. The panel shows that while 73.39% of the records show the strongest gravity accelerations along the y-axis (18% for “-y” and 55.40% for “+y”), suggesting that the phone was in the pocket, we still observe 10.91% of the recordings closer to “-z” (suggesting that the participant was holding the phone in the hand with the screen pointing to the floor), while 15.51% of participants were probably holding the phone in the hand with the phone screen either facing the participant’s leg (“-x”) or facing the opposite side (“x”). Please, note that while it is possible that phones in a pocket can still record strongest gravity accelerations in the x-axis (e.g., when the phone is laying down on its side inside a loose pocket), or that the strongest gravity acceleration is in the y-axis not because the phone is in the pocket, but because the participant is holding it in an upright position, we expect that these cases will be uncommon.

2 Supplementary Note 1

Here, we provide a description of the causality-based confounding tests proposed in reference [1], and illustrate its application to the education, gender, and age variables in the mPower data.

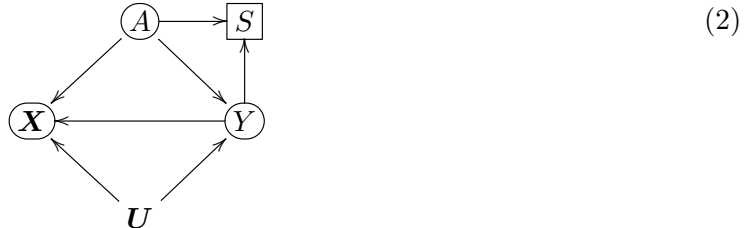
Following Pearl [2], we adopt a mechanism-based approach to causation, where the statistical associations encoded in the joint probability distribution of a set of random variables is supplemented by a directed acyclic graph (DAG) describing the causal relations between the variables. In a causal DAG, a directed edge (arrow) $A \rightarrow B$ indicates that the random variable A is a cause of the random variable B . Under this framework, a variable A is a confounder of the relationship between a feature, X , and the labels, Y , if there is a path from A to X that does not go thorough Y , and there is a path from A to Y that does not go thorough X . As an example, consider the causal diagrams M_1 , M_2 , M_3 , and M_4 presented below.



The variable A is a confounder only on diagram M_1 , since the path $A \rightarrow X$ connects A to X without going through Y , and the path $A \rightarrow Y$ connects A to Y without going through X . For diagram M_2 , the only path connecting A to X goes thorough Y . For diagram M_3 , the only path connecting A to Y goes thorough X . For diagram M_4 there are no paths connecting A to X or A to Y . (Here, we define a path as any unbroken, non-intersecting sequence of directed edges, which may go along or against the direction of the arrows.)

Following the methodology described in reference[1], we employ causality-based tests to evaluate if the predictive performance of a classifier is influenced by measured confounders. The basic idea is to consider the causal diagram describing the data generation process behind the classification task and evaluate if the conditional independence (CI) relationships observed in the data agree with the CI relations implied by the causal diagram. We adopt the notation $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ to describe statistical independence and dependence, respectively (and $A \perp\!\!\!\perp B \mid C$, to describe that A is independent of B conditional on C). Supplementary Figure 11 summarizes the approach, which is briefly described below.

In mobile health studies, the data generation process underlying the data can usually be described by the causal diagram,

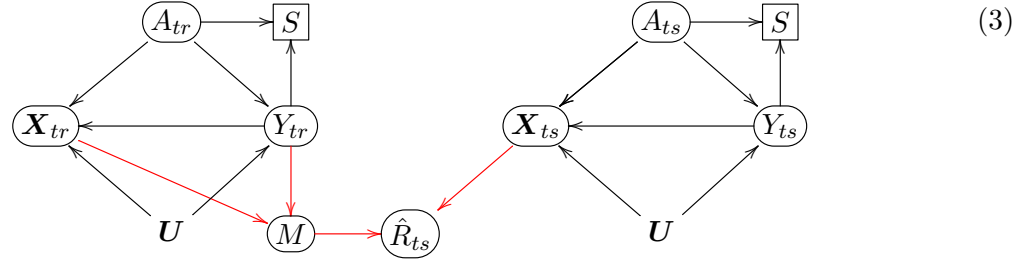


where Y , $\mathbf{X}$, A , $\mathbf{U}$, and S represent, respectively, the disease labels, the set of sensor-based features, an observed demographic confounder, the set of unobserved confounders, and a selection variable (which will be described in more detail below). Note that in diagnostic classification tasks based on sensor features, it is the disease, Y , that has a causal effect on the sensor features, $\mathbf{X}$, since these measurements only capture “symptoms” of the disease. Observe, as well, that for

demographic confounders such as age and gender, it is possible that the confounder influences the features and the disease (but not the other way around).

For concreteness, suppose that, Y represents Parkinson’s disease (PD) status, $\mathbf{X}$ represent a set of features extracted from the accelerometer sensor, A represents age, and U represent any other unobserved confounders that might influence both $\mathbf{X}$ and Y . The arrow from Y to $\mathbf{X}$ indicates that people with PD tend to experience difficulty to walk, so that their acceleration patterns tend to be distinct from control/healthy subjects. The arrow from A to $\mathbf{X}$ shows that age can also potentially influence walking patterns, since old subjects tend to move slower than younger ones. Finally, the arrow from A to Y indicates that age is a risk factor for PD. Furthermore, we assume that the association between A and Y are also influenced by a selection mechanism, represented by the variable S . Here, S represents a binary selection variable indicating whether a potential participant has enrolled or not in the study, that is, $S = 1$ if the person enrolled, and $S = 0$ otherwise. Note that the arrow from A to S indicates that age influences enrollment in the study, since younger people tend to be more technology savvy than older people, and tech savvy people are more likely to enroll in mobile health studies. Observe, as well, that disease status influences enrollment, i.e., $Y \rightarrow S$, since subjects suffering from a disease are usually more motivated to enroll in a study than controls. The squared frame around S in panel b indicates that the analysis is conditional on the participants that actually enrolled in the study, that is, the analysis is conditional on $S = 1$.

The causal diagram,



describes the data generation process underlying a digital health classification task based on sensor features. The subscripts tr and ts indicate the training and test sets, respectively, M represents the trained classification model, and $\hat{R}_{ts}$ represents the predicted positive class probability generated by the classifier. Note that $\hat{R}_{ts}$ is a function of the training data (i.e., training labels, Y_{tr} , and training features, $\mathbf{X}_{tr}$), as well as, the test features, $\mathbf{X}_{ts}$, since $\hat{R}_{ts}$ is generated by evaluating the test features using the model trained on the training data.

The causal diagram,



represents a simplified version of the classification task diagram, focusing only on the prediction scores, test set labels and confounders, where the path $A_{ts} \rightarrow \mathbf{X}_{ts} \rightarrow \hat{R}_{ts}$ is replaced by the collapsed version $A_{ts} \rightarrow \hat{R}_{ts}$, and the paths $Y_{ts} \rightarrow \mathbf{X}_{ts} \rightarrow \hat{R}_{ts}$ and $Y_{ts} \leftarrow U \rightarrow \mathbf{X}_{ts} \rightarrow \hat{R}_{ts}$ are replaced by the path $Y_{ts} \rightarrow \hat{R}_{ts}$. Application of the the d-separation criterion [2] indicates that

the CI relations in the the first row of Table 1 (CI pattern 1) should hold in the data, if A is really confounding the prediction. (Here, we are assuming that the faithfulness assumption [4, 2] holds in the data.)

Therefore, in order to determine if the observed variable A is indeed a confounder, we can simply check if the CI relations in the data match the CI pattern in the first row of Table 1. For instance, if A does not have a direct causal effect on $\mathbf{X}$ (such as in diagram M_2 in (1)), then the simplified classification task diagram will be given by,



and the CI relations will be given by the second row of Table 1 (CI pattern 2). Similarly, if A does not have a direct causal effect on Y , and there are no selection biases (such as in diagram M_3 in (1)) then the simplified classification task graph will be given by,



and the CI relations will be given by the third row of Table 1 (CI pattern 3). Finally, if A is not connected to $\mathbf{X}$ or Y (such as in diagram M_4 in (1)), then the simplified classification task graph will correspond to,



and the CI relations will be given by the fourth row of Table 1 (CI pattern 4).

Hence, by comparing the CI patterns in the data to the CI patterns in Table 1, we can determine whether A is confounding the prediction (pattern 1) or not (patterns 2, 3, and 4).

Observe, as well, that these tests can also be used to investigate if a confounding adjustment is working as expected. For instance, after the matching adjustment, we would expect A and Y to be marginally independent, so that the CI pattern after the matching adjustment should correspond to either pattern 3 or 4. Supplementary Figure 12 presents the simplified graphs alongside the associated CI patterns described in Table 1.

In this paper, we performed two sets of confounding analyses. In the first, we evaluated whether age, gender, and education were confounding the predictions from separated PD vs non-PD classifiers trained with tapping, walk, balance, and voice features. (We selected these 3 variables because they were associated with PD status and, therefore, could potentially confound the analyses.) The goal of these first analyses was to determine which variables were, indeed, confounding the predictions, so that we could only use the detected confounders for matching the data. (Note that the larger the number of confounder variables, the harder it is to achieve

CI patterns	Conditional (in)dependence relations
CI pattern 1	$\hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp A_{ts} , A_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} \mid A_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp A_{ts} \mid Y_{ts}$
CI pattern 2	$\hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp A_{ts} , A_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} \mid A_{ts} , \hat{R}_{ts} \perp\!\!\!\perp A_{ts} \mid Y_{ts}$
CI pattern 3	$\hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp A_{ts} , A_{ts} \perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} \mid A_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp A_{ts} \mid Y_{ts}$
CI pattern 4	$\hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \perp\!\!\!\perp A_{ts} , A_{ts} \perp\!\!\!\perp Y_{ts} , \hat{R}_{ts} \not\perp\!\!\!\perp Y_{ts} \mid A_{ts} , \hat{R}_{ts} \perp\!\!\!\perp A_{ts} \mid Y_{ts}$

Table 1: Conditional independence patterns used to determine if a variable A is confounding the predictive performance of a classifier. Here, A_{ts} , Y_{ts} , and $\hat{R}_{ts}$ represent, respectively, the potential confounder, the label, and the predicted class probability in the test set. Only pattern 1 indicates the presence of confounding.

balance across all variables during matching. Hence, determining which variables are really confounding the prediction is important in practice, as matching on unnecessary variables often leads to unnecessary additional reductions in sample size.) Because only age (and gender to a lesser degree) were flagged as confounders, we only performed age/gender matching. In our second set of analyses, we reanalyzed the data after matching for age and gender in order to make sure that the predictions based on the matched data were not confounded.

In all analyses we employed correlation tests (using the `cor.test` function in R) and partial correlation tests (using the `ppcor` R package) to test for marginal and conditional associations, respectively. In all figures, the boxplots represent the distribution of the results (correlations and respective corrected p-values) across 100 random splits of the data into training and test sets (with half of the data used for training and the remaining half for testing). We performed multiple testing correction across the 100 data splits using the Benjamini-Hochberg procedure [5]. For the matched data we also performed distance correlation tests [6] and partial distance correlation tests [7] (implemented in the `energy` R package) to make sure the (partial) correlation tests were not missing non-linear associations. For all analyses, we binarized the gender variable, setting “Female = 1” and “Male = 0”, and we recoded the education variable using a numeric scale ranging from 1 to 7 (coding the original ordinal levels as “Some high school = 1”, “School Diploma/GED = 2”, “Some college = 3”, “2-year college degree = 3”, “4-year college degree = 4”, “Some graduate school = 5”, “Master’s Degree = 6”, “Doctoral degree = 7”).

Supplementary Figures 13, 14, and 15 present the random forest results for the age, gender, and education variables, respectively, while Supplementary Figures 16, 17, and 18 present analogous analyses based on a ridge-regression classification, for the first set of analyses based on matched data.

In all these figures, the top row panels show correlations and partial correlations among the labels (Y), confounder (A), and classifier predictions (R) for each of the 4 activity assessments. The bottom row panels show the multiple testing adjusted p-values (in $-\log_{10}$ scale) for the respective correlation and partial correlation tests. The red horizontal line is set at $-\log_{10}(0.05)$, so that results above the red line are considered statistically significant at a 0.05 level. Situations where the boxplots are entirely above the red line show evidence against the null hypothesis of zero correlation. Situations where the boxplots are only partially above the red line (but where the median is still above the red line) represent only marginal evidence (since the results are

statistically significant for most but not all data splits). To determine if a variable is confounding the prediction we compare the patterns of conditional independence relations observed in the data against the patterns presented in Supplementary Figure 12¹ and Table 1.

Supplementary Figures 13 and 16 show clear evidence that age is confounding the predictions for the tapping and walk assessments, and marginal evidence for the balance and voice assessments. (Note how the conditional independence patterns match CI pattern 1 in the first row of Table 1 and Figure 12a.) Supplementary Figures 14 and 17 show some evidence that gender might be confounding the predictions for voice (and tapping and walk to a lesser degree). Supplementary Figures 15 and 18, on the other hand, show that, for the education variable, the conditional independence patterns were mostly consistent with CI pattern 2 (where R and A are conditionally independent given Y , as shown in the second row of Table 1 and Supplementary Figure 12b). These patterns suggest that education is not confounding the prediction. Overall, because the analyses showed that only age (and gender to a considerably lesser degree) were confounding the predictions, we only needed to perform age/gender matching.

In the second set of analyses, we reanalyzed the age/gender matched data in order to make sure that our predictions were no longer confounded. Supplementary Figures 19, 20, and 21 present the results for random forest classifiers on the matched data, while Supplementary Figures 22, 23, and 24 present analogous results based on ridge-regression classification. As expected, the confounding tests no longer found evidence of confounding biases in the predictions showing that the matching adjustment was effective. Note that most of the analyses showed conditional independence (CI) relations consistent with CI patterns 3 or 4 in Supplementary Figure 12, indicating the absence of confounding biases. The exceptions were for ridge-regression classifiers trained with the voice features, where gender (Supplementary Figure 23h) showed CI relations that were consistent with the model,

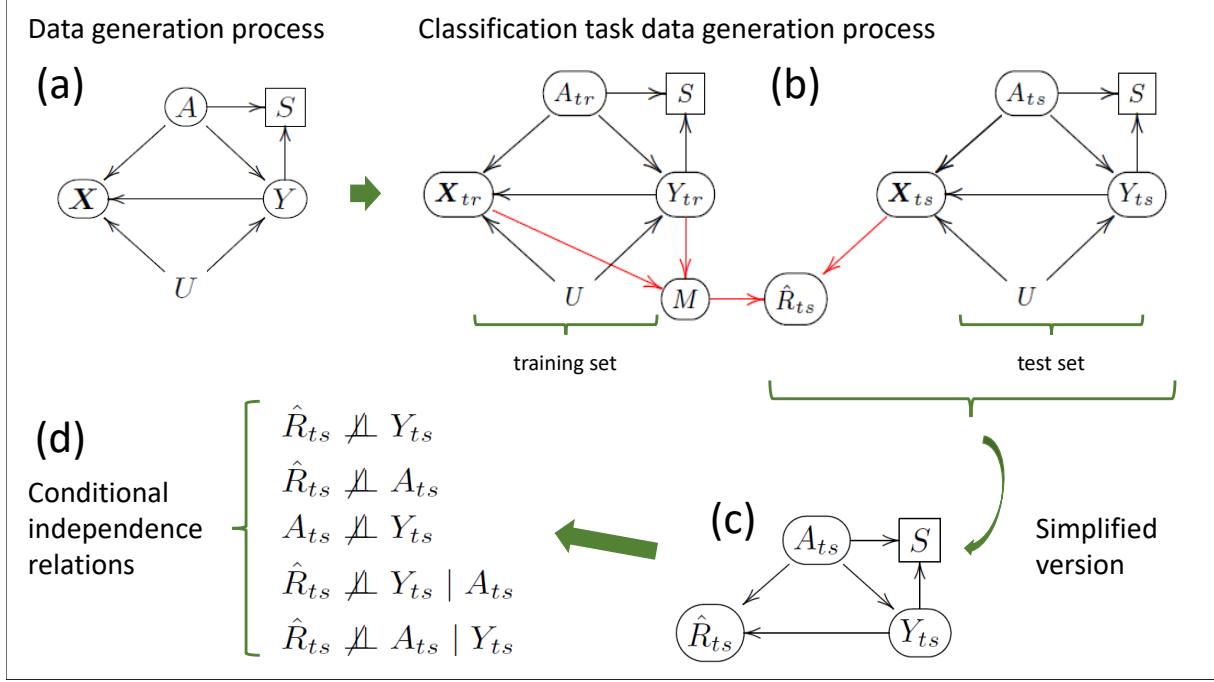


and education (Supplementary Figure 24h) showed CI relations consistent with a model where $A \not\rightarrow R$, $A \not\rightarrow Y$, and $Y \not\rightarrow R$. (Here, the notation $A \not\rightarrow B$ means that there is no arrow from A to B .)

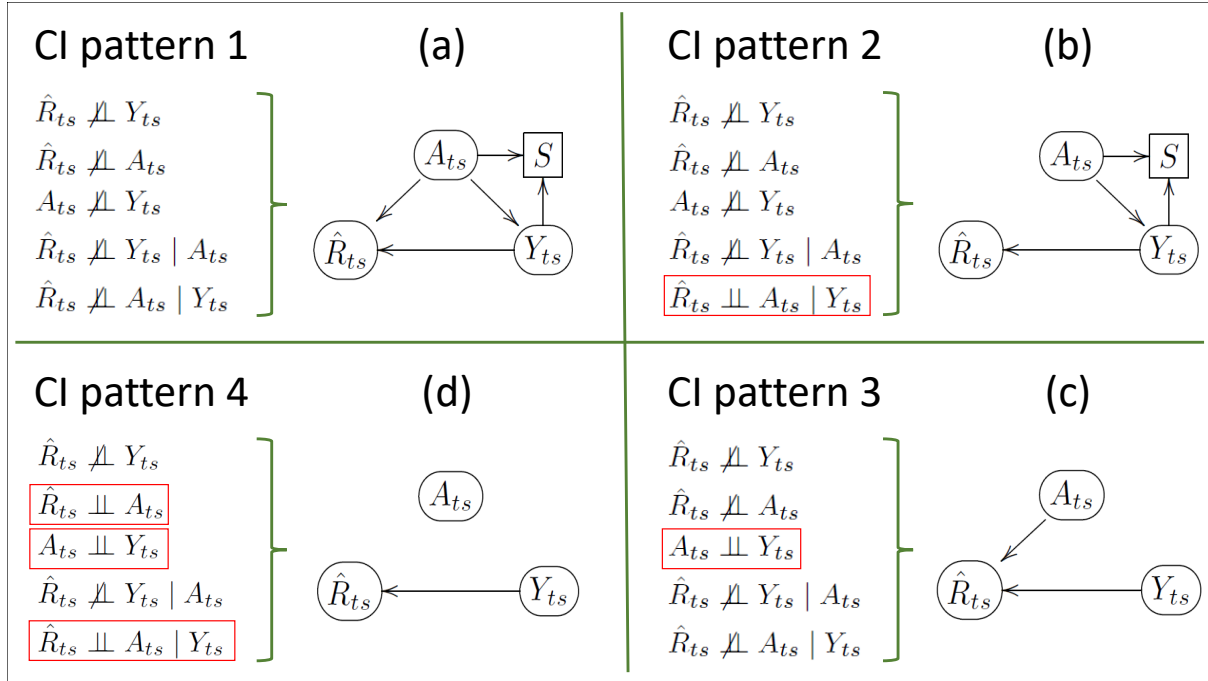
For the matched data we also performed distance correlation tests [6] and partial distance correlation tests [7] (implemented in the **energy** R package) to make sure the (partial) correlation tests were not missing non-linear associations. Supplementary Figures 25 to 28 compare the p-values derived from the correlation tests and distance correlation tests. The results of the (partial) distance correlation tests and standard (partial) correlation tests were for the most part largely consistent. (The exception was for the ridge-regression classifiers based on the voice features, which showed marginally significant non-linear associations for a few data splits.)

¹Note, however, that Supplementary Figure 12 does not provide an exhaustive list of possible patterns, and that it is possible to observe patterns consistent with other models. For instance, Supplementary Figure 23d and h are consistent with a model where $A \not\rightarrow R$, $A \not\rightarrow Y$, and $Y \not\rightarrow R$, which is not shown in Supplementary Figure 12 (here, the notation $A \not\rightarrow B$ means that there is no arrow from A to B). Also, in practice, it is sometimes possible to observe “unfaithful” CI patterns that are not compatible with the assumed causal model.

2.1 Supplementary Figures 11 to 30

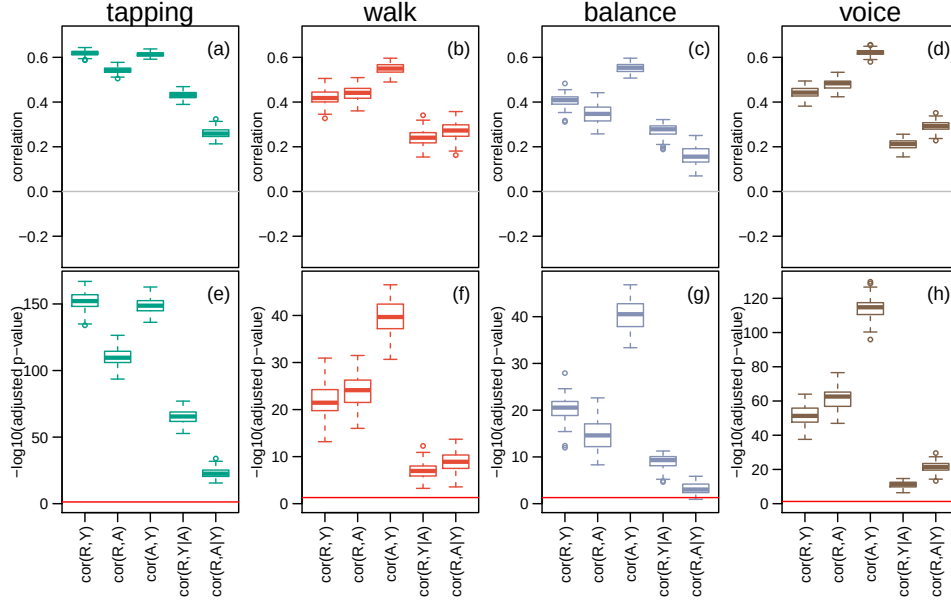


Supplementary Figure 11: Overview of the causality based approach to detect the influence of observed confounders. Panel a shows the causal diagram for data collected by a sensor based mobile-health study, where Y , $\mathbf{X}$, A , and $\mathbf{U}$ represent, respectively, the disease labels (Y), the sensor-based features ($\mathbf{X}$), the observed confounder (A), and the unobserved confounders ($\mathbf{U}$). Panel b shows the data generation process for the associated digital health classification task, where the subscripts tr and ts indicate the training and test sets, respectively, M represents the trained model, and $\hat{R}_{ts}$ represents the predicted positive class probability generated by the classifier. Panel c shows the simplified version of the classification task, focusing only on $\hat{R}_{ts}$, A_{ts} , and Y_{ts} . Finally, panel d shows the conditional independence relations implied by the diagram on panel c. We can check if these conditional and marginal associations hold in the data in order to determine if A is confounding the prediction.



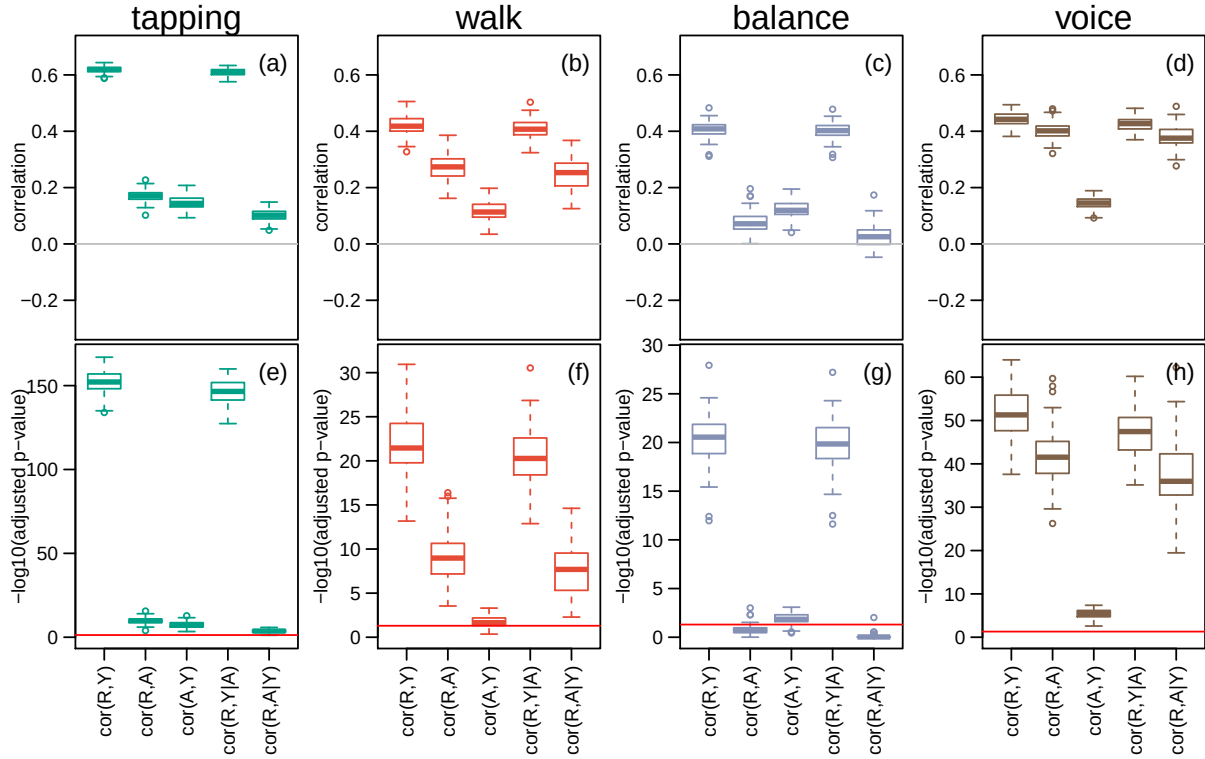
Supplementary Figure 12: Conditional independence patterns used to assess if the variable A is confounding the predictive performance of a classifier. A_{ts} , Y_{ts} , and $\hat{R}_{ts}$ represent, respectively, the potential confounder, the label, and the predicted class probability in the test set. CI pattern 1 (panel a) indicates that A is a confounder. The other CI patterns in panels b, c, and d, indicate that A is not confounding the predictive performance.

age – unmatched – random forest



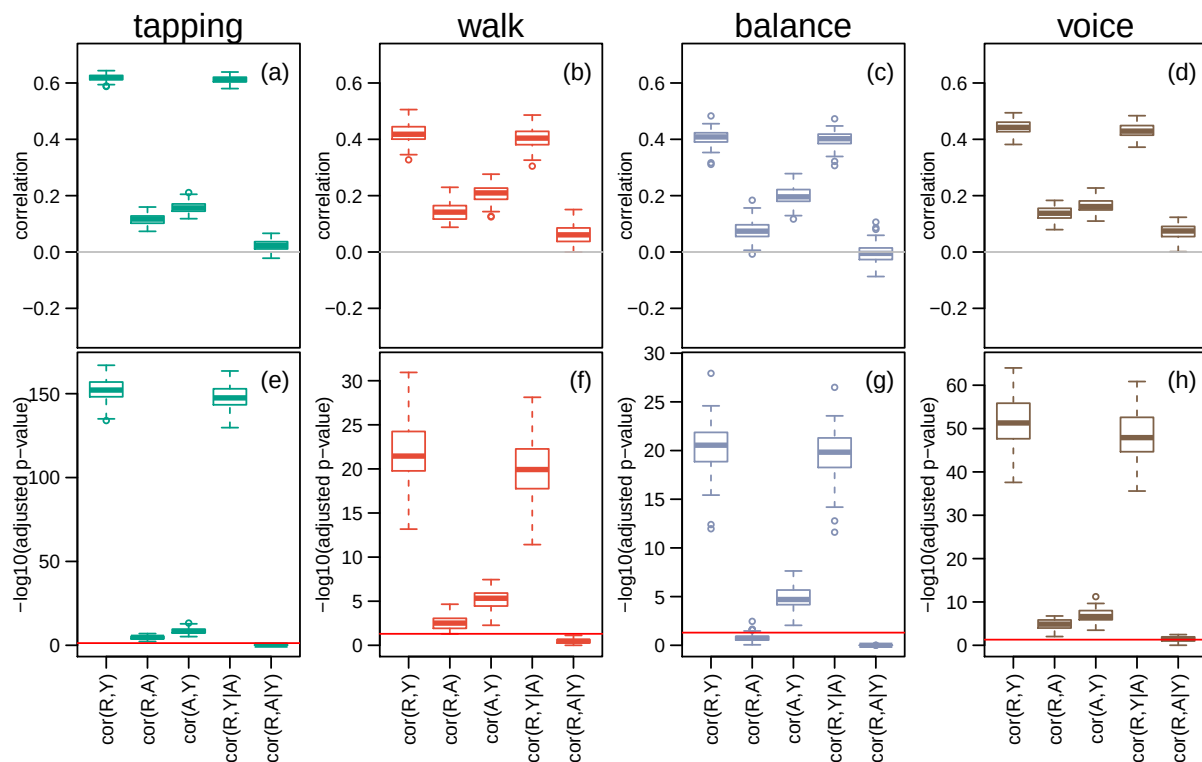
Supplementary Figure 13: Checking for age confounding in the predictions of a random forest classifier trained with unmatched data. The top row panels show correlations and partial correlations among the labels (Y), confounder (A), and classifier predictions (R) for each activity assessments. The bottom row panels show the multiple testing adjusted p-values (in $-\log_{10}$ scale) for the respective correlation and partial correlation tests. The boxplots represent the results from 100 random splits of the data into training and test sets. The red horizontal line (set at $-\log_{10}(0.05)$) corresponds to a 0.05 statistical significance threshold. (For each boxplot, the box corresponds to the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.) Situations where the boxplots are entirely above the red line show evidence against the null hypothesis of zero (partial) correlation. Situations where the boxplots are only partially above the red line (but where the median is still above the red line) represent marginal evidence (since the results are statistically significant for most but not all data splits). To determine if a variable is confounding the prediction we compare the patterns of conditional independence relations observed in the data against the patterns presented in Supplementary Figure 12. The results are clearly consistent with the CI pattern 1 (Supplementary Figure 12a) for the tapping, walk, and voice assessments (and marginally consistent for the balance assessment) suggesting that age is an important confounder of the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362$, 1,092, 1,099, and 2,410, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction.

gender – unmatched – random forest



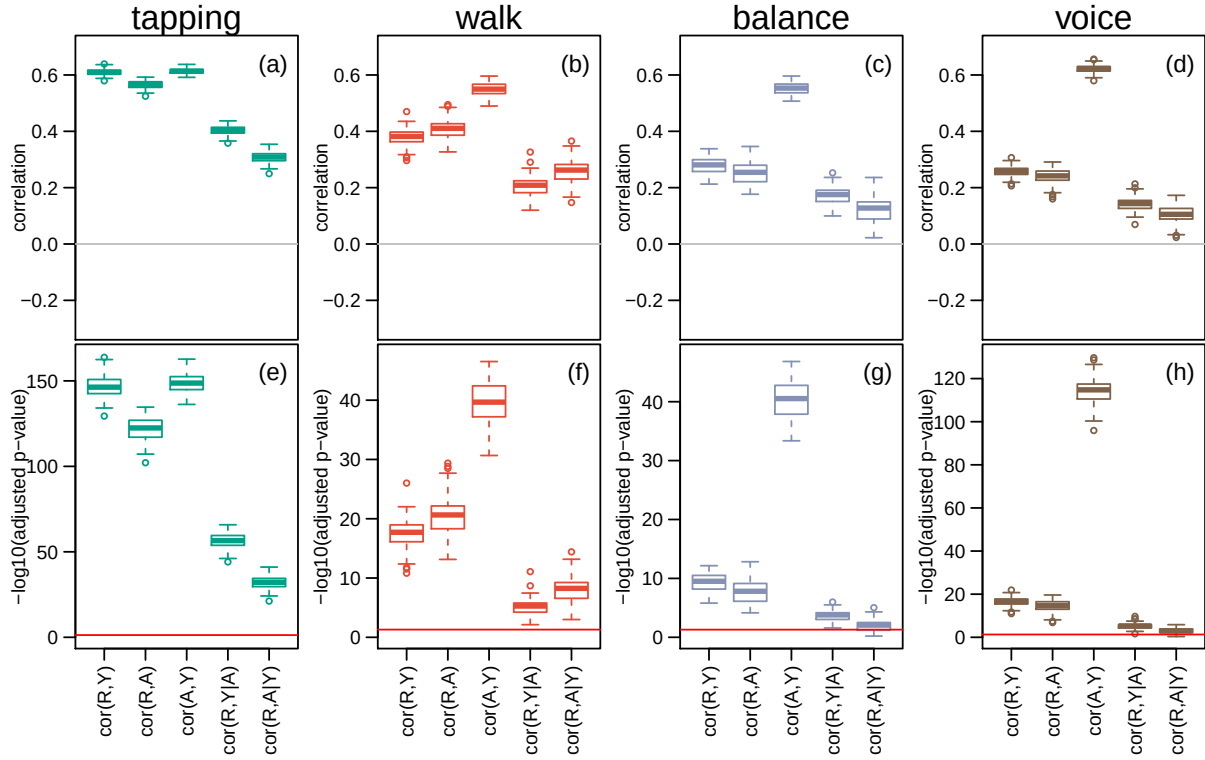
Supplementary Figure 14: Checking for gender confounding in the predictions of a random forest classifier trained with unmatched data. The results are consistent with the CI pattern 1 (Supplementary Figure 12a) for the voice assessment, and marginally consistent with CI patterns 2, 3, and 4 for the tapping, walk, and balance assessments, respectively. Overall, these results suggest that gender has only a marginal influence in the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362, 1,092, 1,099,$ and $2,410$, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – unmatched – random forest



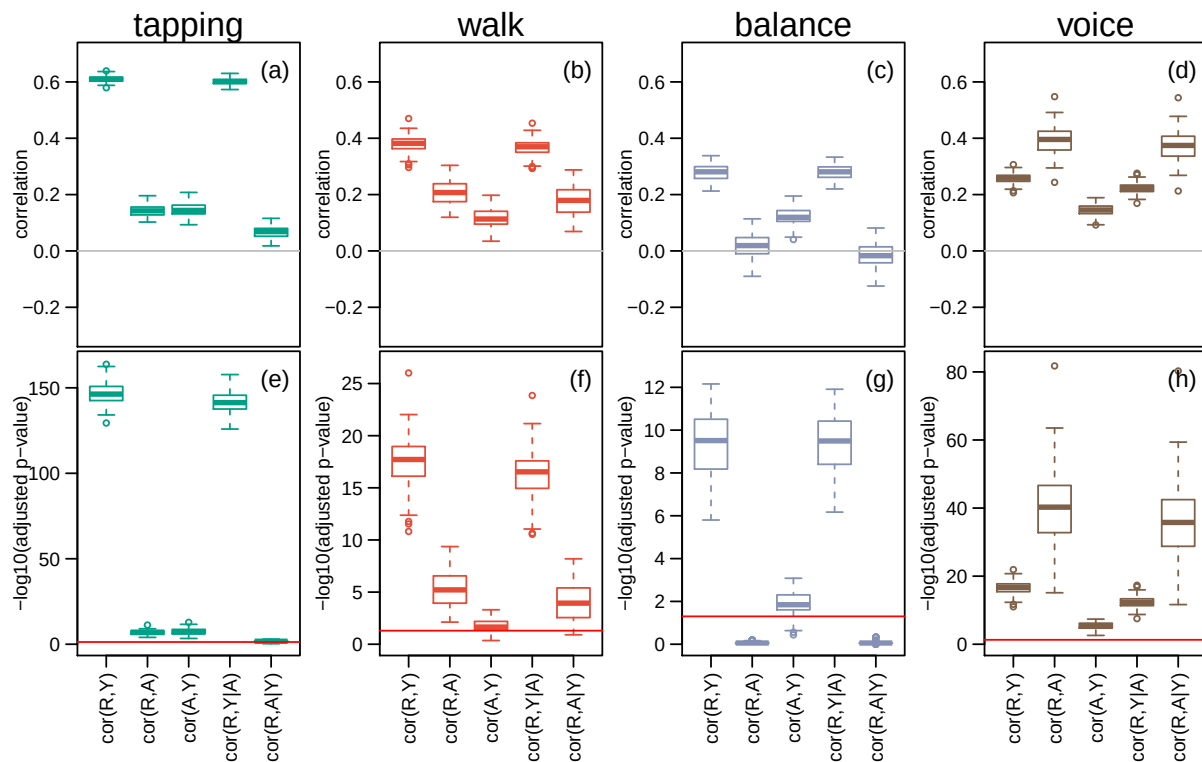
Supplementary Figure 15: Checking for education confounding in the predictions of a random forest classifier trained with unmatched data. The results are consistent with the CI pattern 2 (Supplementary Figure 12b) for the tapping, walk, and voice assessments, suggesting that education is not a confounder of the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362, 1,092, 1,099$, and $2,410$, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

age – unmatched – ridge regression



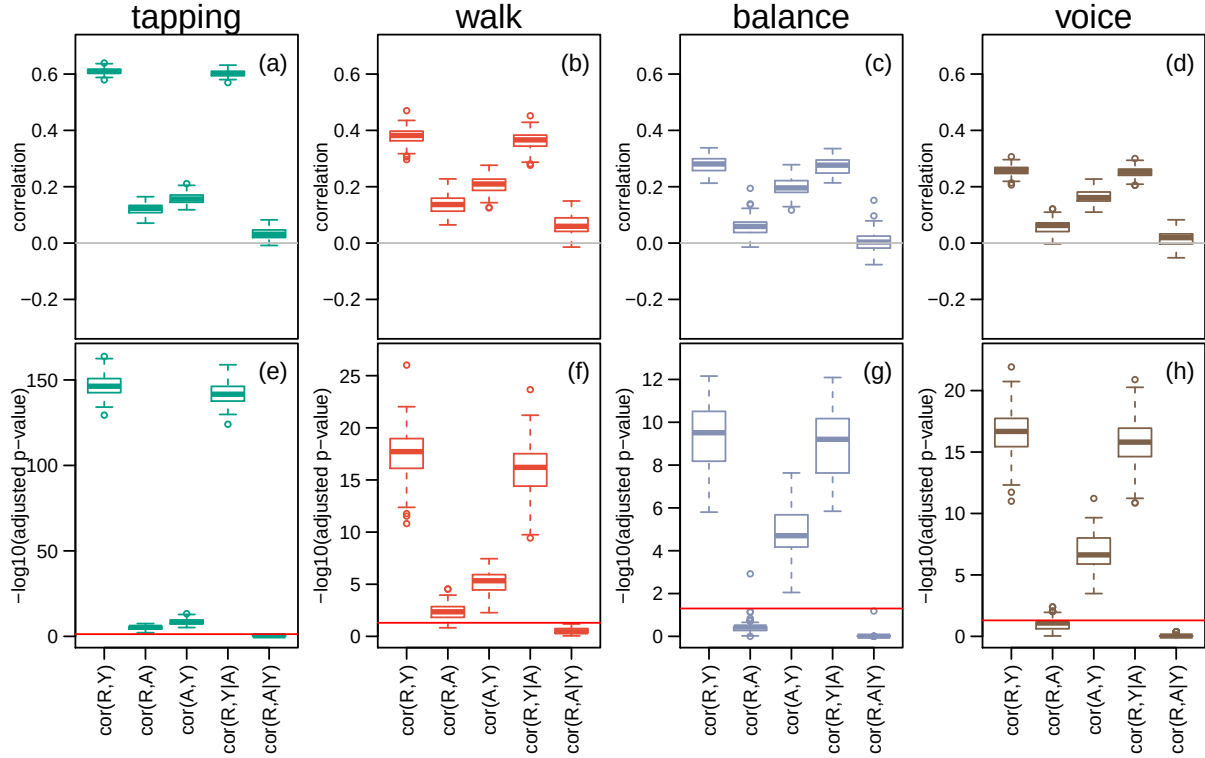
Supplementary Figure 16: Checking for age confounding in the predictions of a ridge-regression classifier trained with unmatched data. The results are clearly consistent with the CI pattern 1 (Supplementary Figure 12a) for the tapping and walk assessments (and marginally consistent for the balance and voice assessments) suggesting that age is an important confounder of the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362, 1,092, 1,099$, and $2,410$, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

gender – unmatched – ridge regression

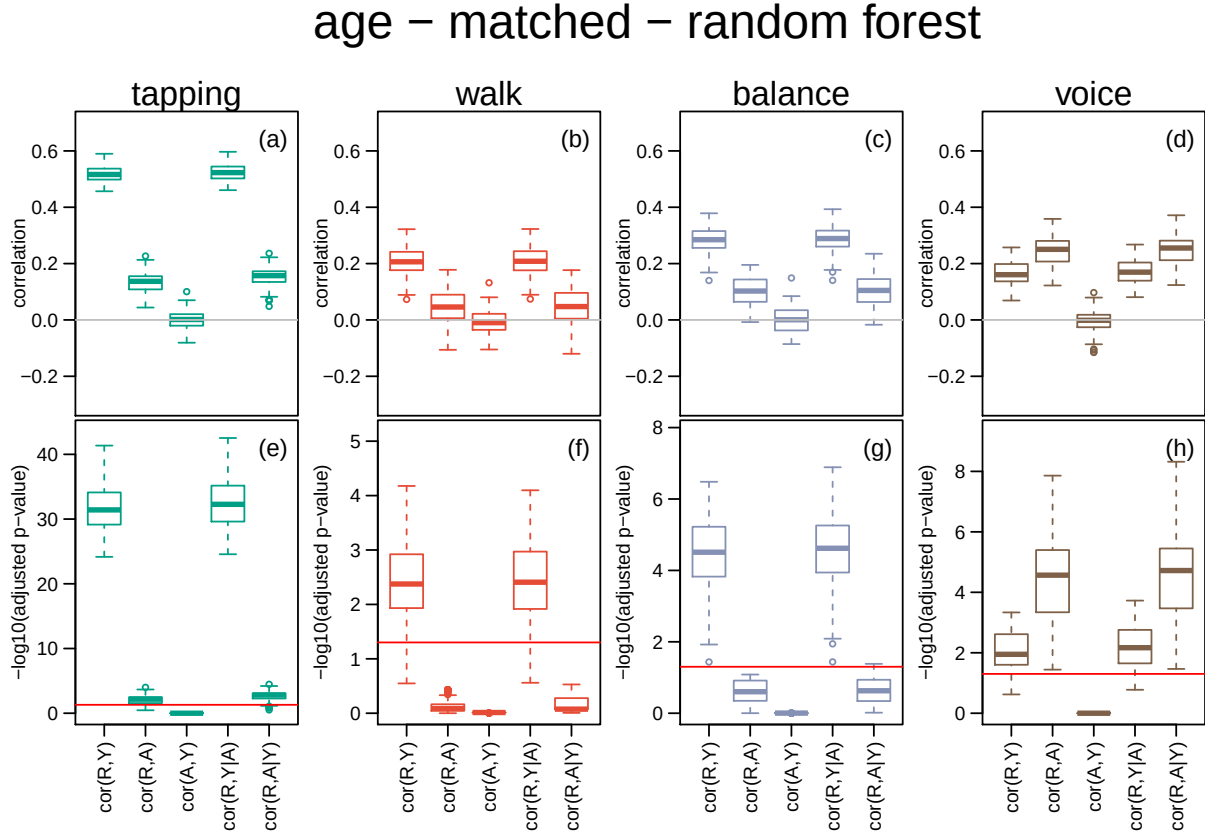


Supplementary Figure 17: Checking for gender confounding in the predictions of a ridge-regression classifier trained with unmatched data. The results are consistent with the CI pattern 1 (Supplementary Figure 12a) for the voice assessment, and marginally consistent with CI patterns 2, 3, and 4 for the tapping, walk, and balance assessments, respectively. Overall, these results suggest that gender has only a marginal influence in the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362, 1,092, 1,099,$ and $2,410$, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – unmatched – ridge regression

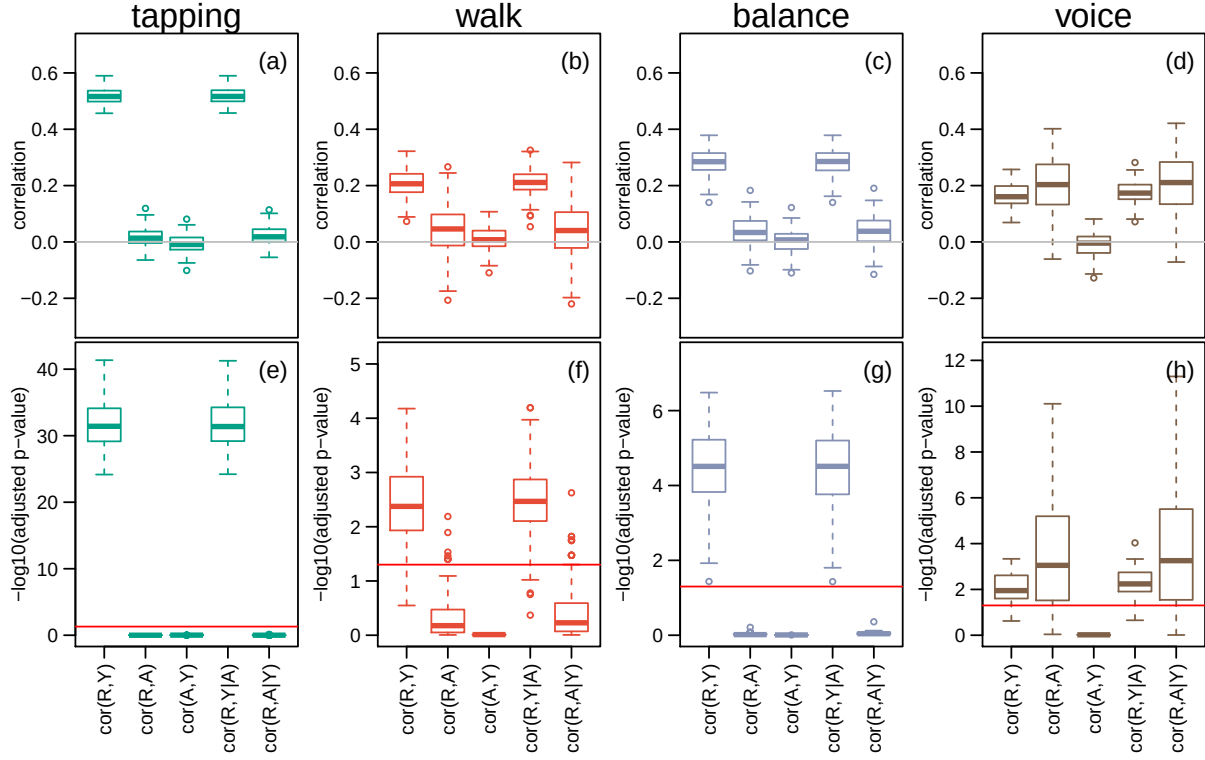


Supplementary Figure 18: Checking for education confounding in the predictions of a ridge-regression classifier trained with unmatched data. The results are mostly consistent with CI pattern 2 (Supplementary Figure 12b) suggesting that education is not a confounder of the predictions. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 3,362, 1,092, 1,099$, and $2,410$, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.



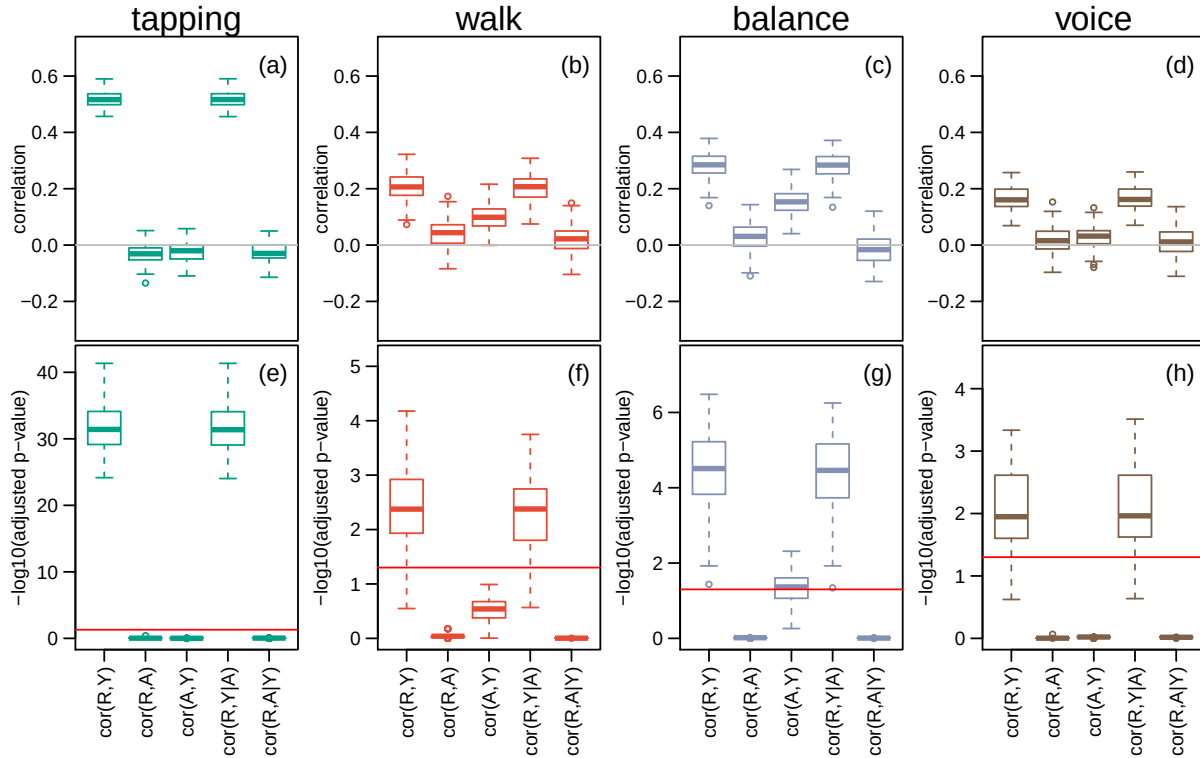
Supplementary Figure 19: Checking for age confounding in the predictions of a random forest classifier trained with matched data. The results are consistent with the CI pattern 3 (Supplementary Figure 12c) for the tapping and voice assessments, and consistent with CI pattern 4 (Supplementary Figure 12d) for the walk and balance assessments. These results suggest that age is not confounding the predictions in the matched data. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960, 464, 466$, and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

gender – matched – random forest



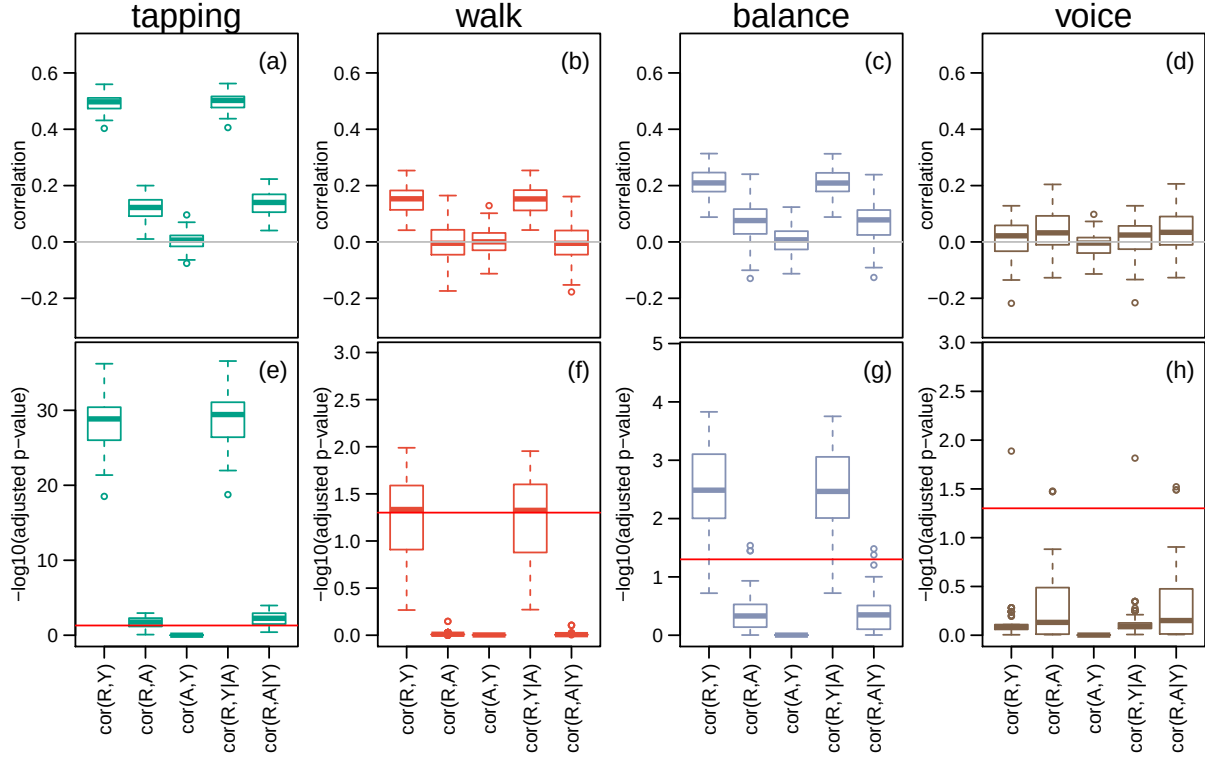
Supplementary Figure 20: Checking for gender confounding in the predictions of a random forest classifier trained with matched data. The results are consistent with the CI pattern 3 (Supplementary Figure 12c) for the voice assessment, and consistent with CI pattern 4 (Supplementary Figure 12d) for the tapping, walk and balance assessments. These results suggest that gender is not confounding the predictions in the matched data. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960, 464, 466,$ and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – matched – random forest



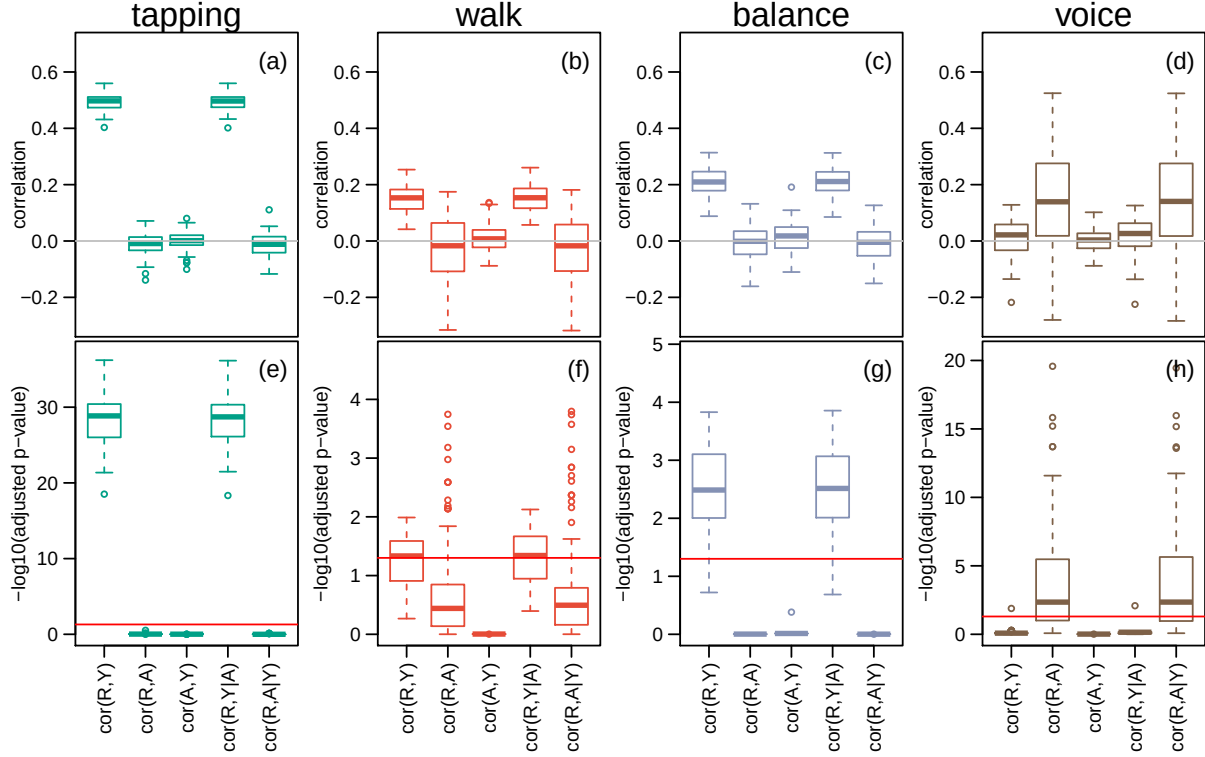
Supplementary Figure 21: Checking for education confounding in the predictions of a random forest classifier trained with matched data. The results are mostly consistent with the CI pattern 4 (Supplementary Figure 12d), suggesting that education is not confounding the predictions in the matched data. P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960, 464, 466,$ and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

age – matched – ridge regression



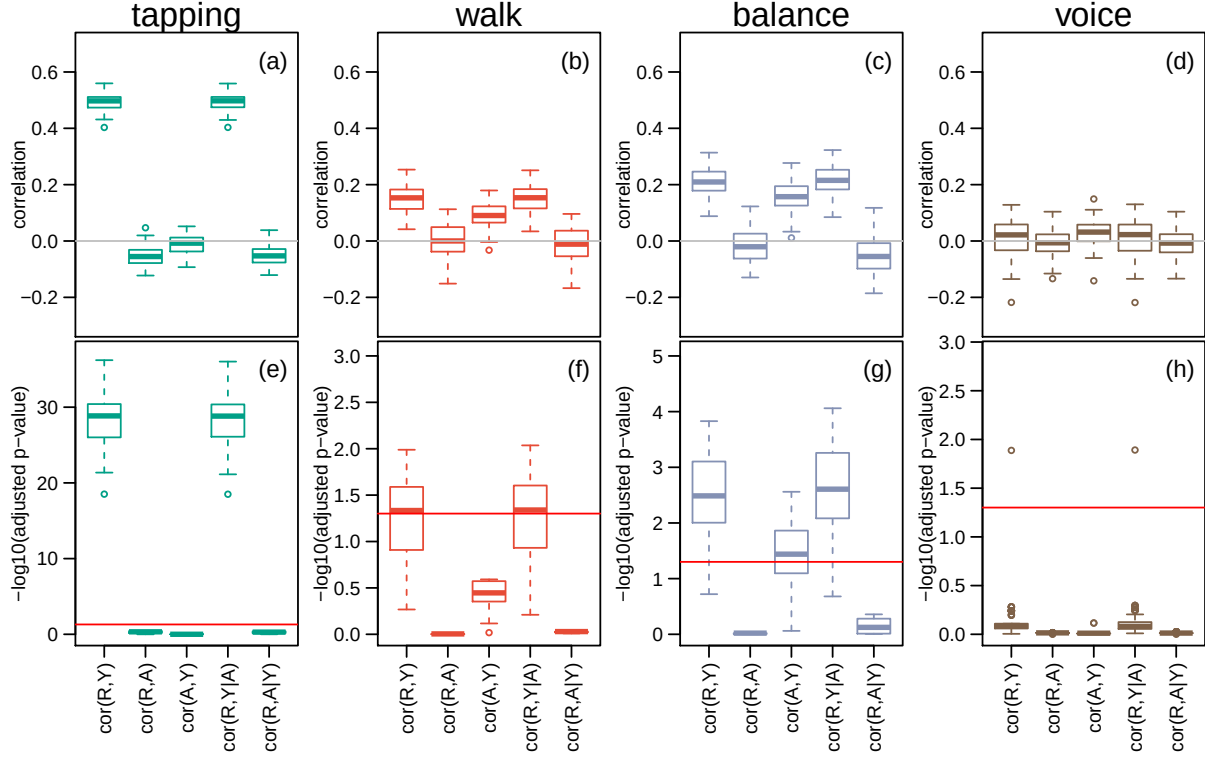
Supplementary Figure 22: Checking for age confounding in the predictions of a ridge-regression classifier trained with matched data. The results suggest that age is not confounding the predictions in the matched data. (The tapping assessment is consistent with CI pattern 2, while the walk and balance assessments are consistent with CI pattern 4. The voice assessment is consistent with a model where $A \not\rightarrow R$, $A \not\rightarrow Y$, and $Y \not\rightarrow R$). P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960, 464, 466$, and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

gender – matched – ridge regression



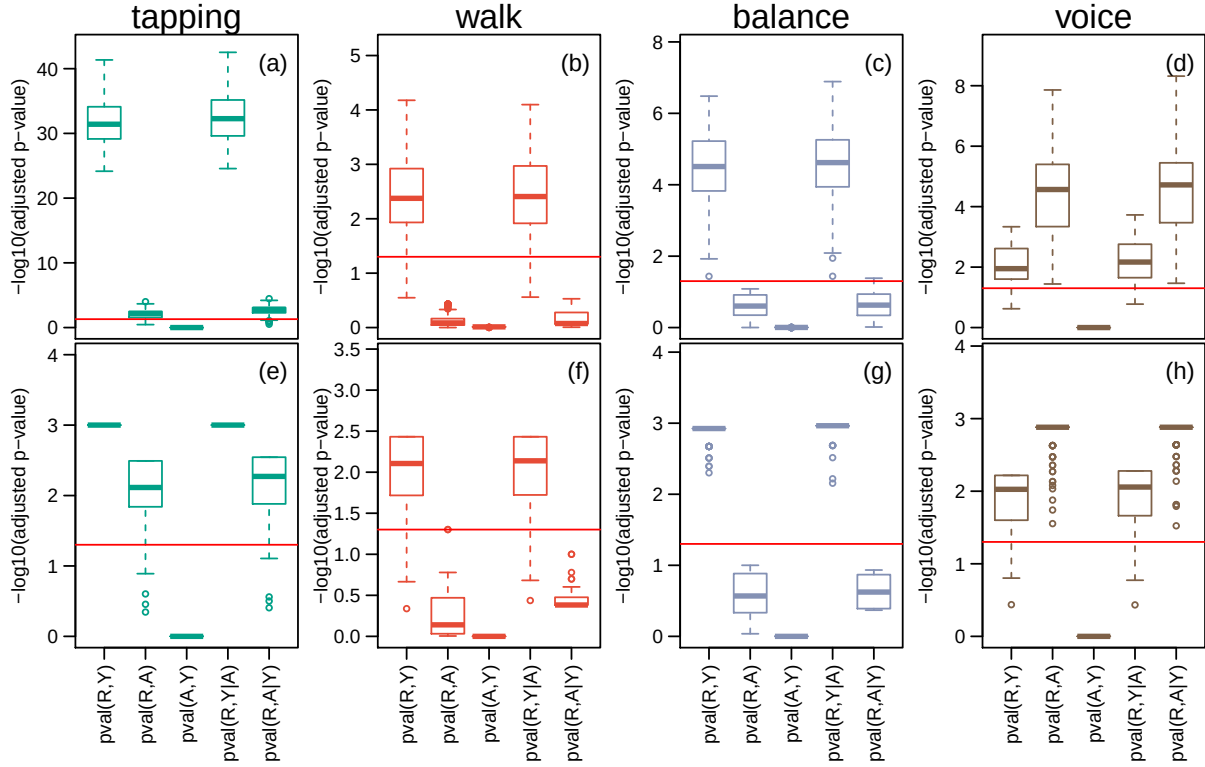
Supplementary Figure 23: Checking for gender confounding in the predictions of a ridge-regression classifier trained with matched data. The results suggest that gender is not confounding the predictions in the matched data. (The tapping, walk, and balance assessments are consistent with CI pattern 4. The voice assessment is consistent with a model where $A \rightarrow R$, $A \not\rightarrow Y$, and $Y \not\rightarrow R$). P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960$, 464, 466, and 620, for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – matched – ridge regression



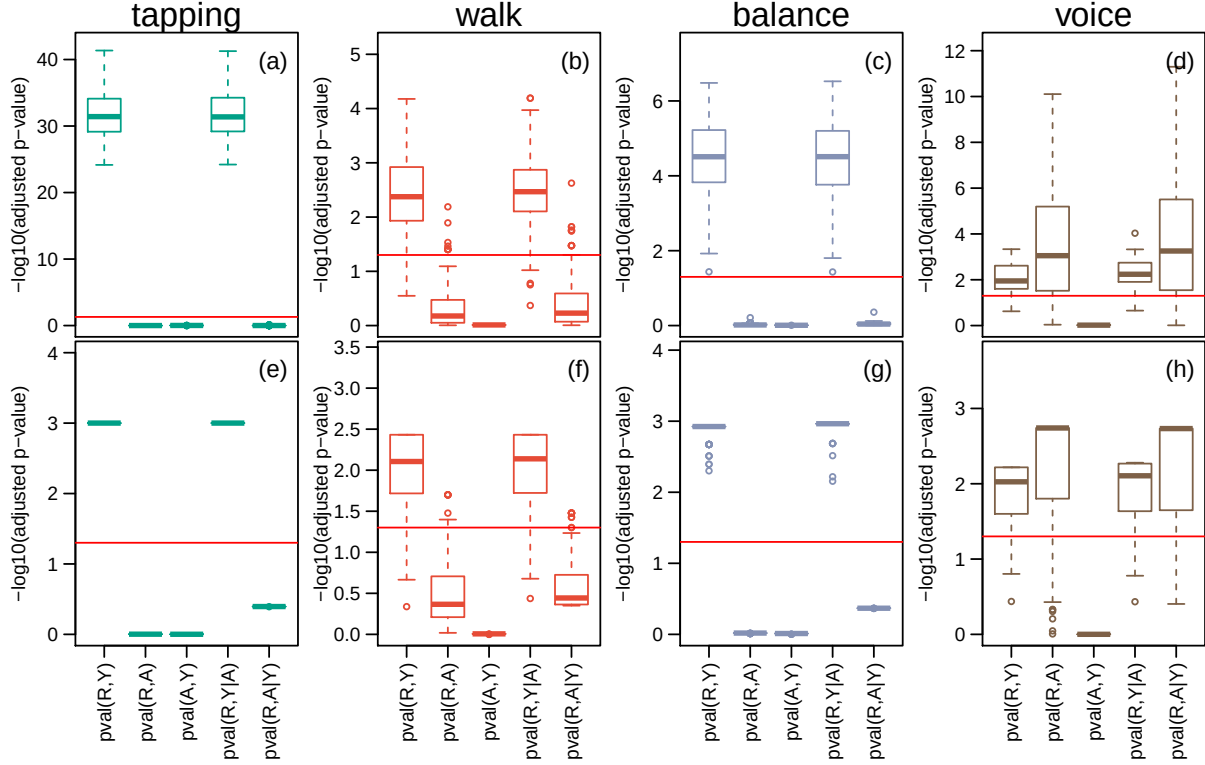
Supplementary Figure 24: Checking for education confounding in the predictions of a ridge-regression classifier trained with matched data. The results suggest that gender is not confounding the predictions in the matched data. (The tapping, walk, and balance assessments are consistent with CI pattern 4. The voice assessment is consistent with a model where $A \not\rightarrow R$, $A \not\rightarrow Y$, and $Y \not\rightarrow R$). P-values were derived from a two-tailed (partial) correlation test (testing the null hypothesis that the (partial) correlation was equal to zero), based on $n = 960$, 464 , 466 , and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

age – cor. vs distance cor. tests – random forest



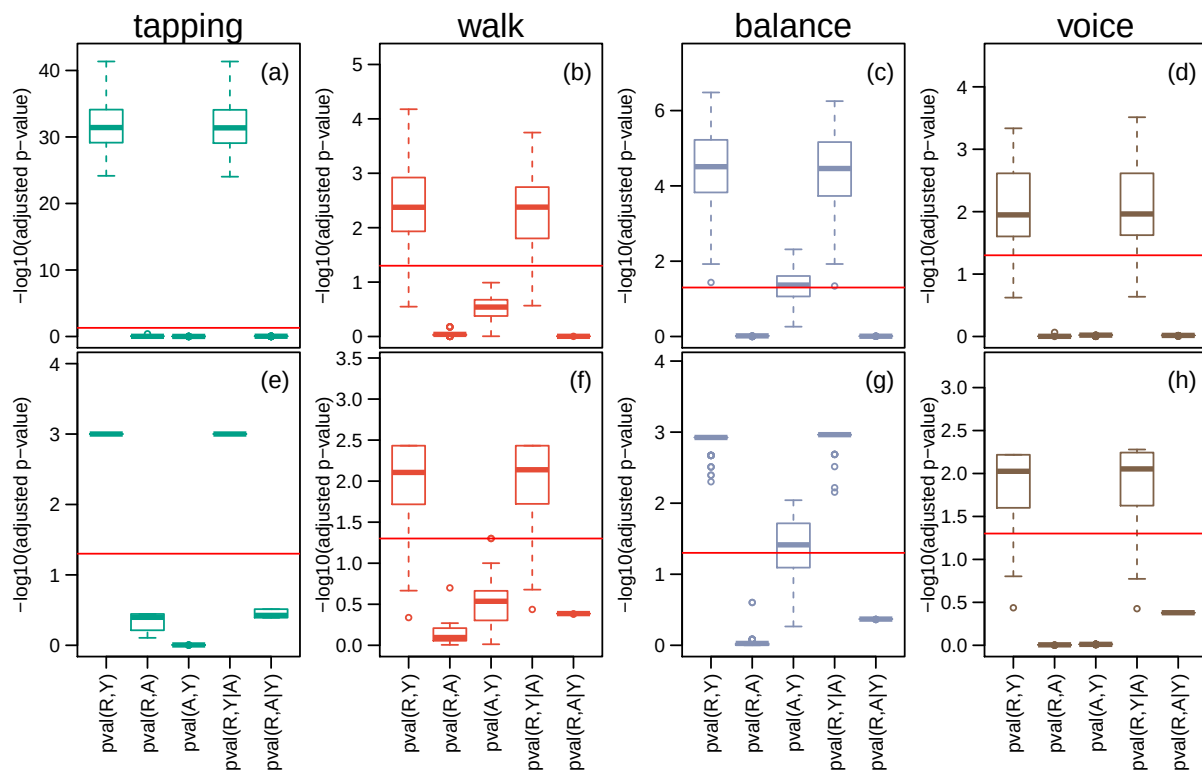
Supplementary Figure 25: Comparison of (partial) correlation and (partial) distance correlation tests for age confounding in the predictions of a random forest classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are consistent. Analyses based on $n = 960, 464, 466,$ and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

gender – cor. vs distance cor. tests – random forest



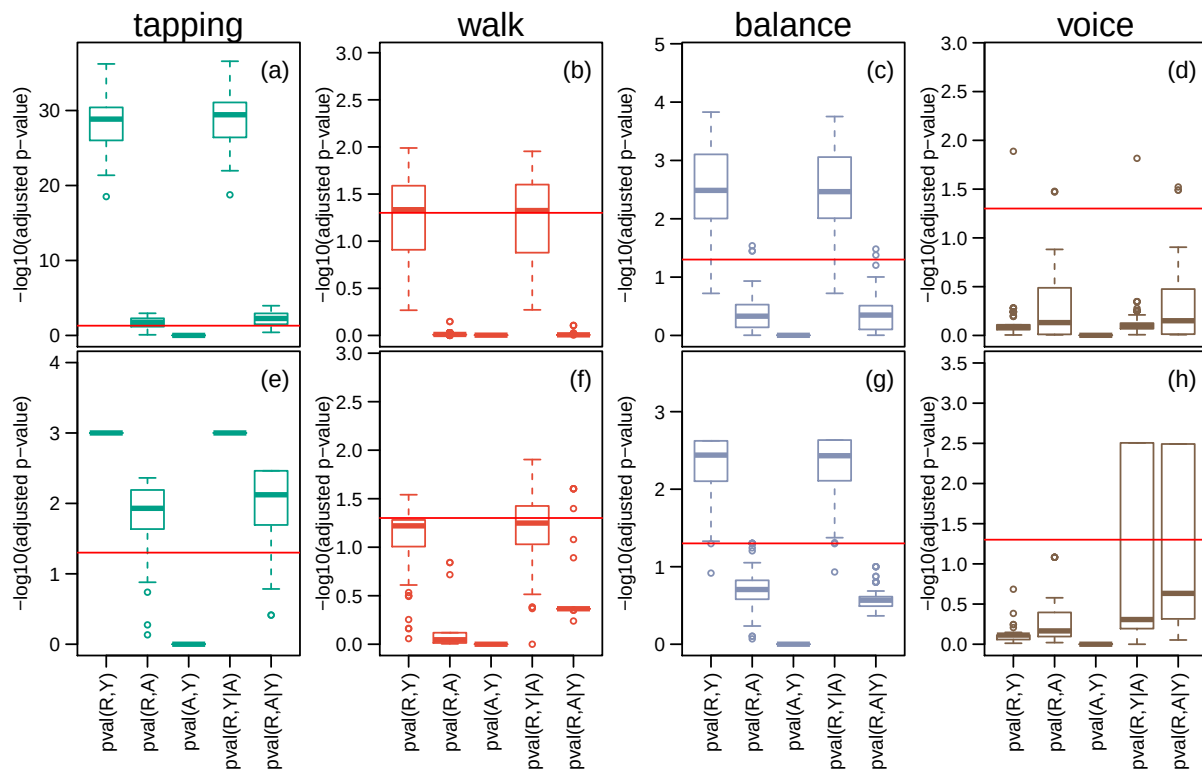
Supplementary Figure 26: Comparison of (partial) correlation and (partial) distance correlation tests for gender confounding in the predictions of a random forest classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are consistent. Analyses based on $n = 960, 464, 466,$ and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – cor. vs distance cor. tests – random forest



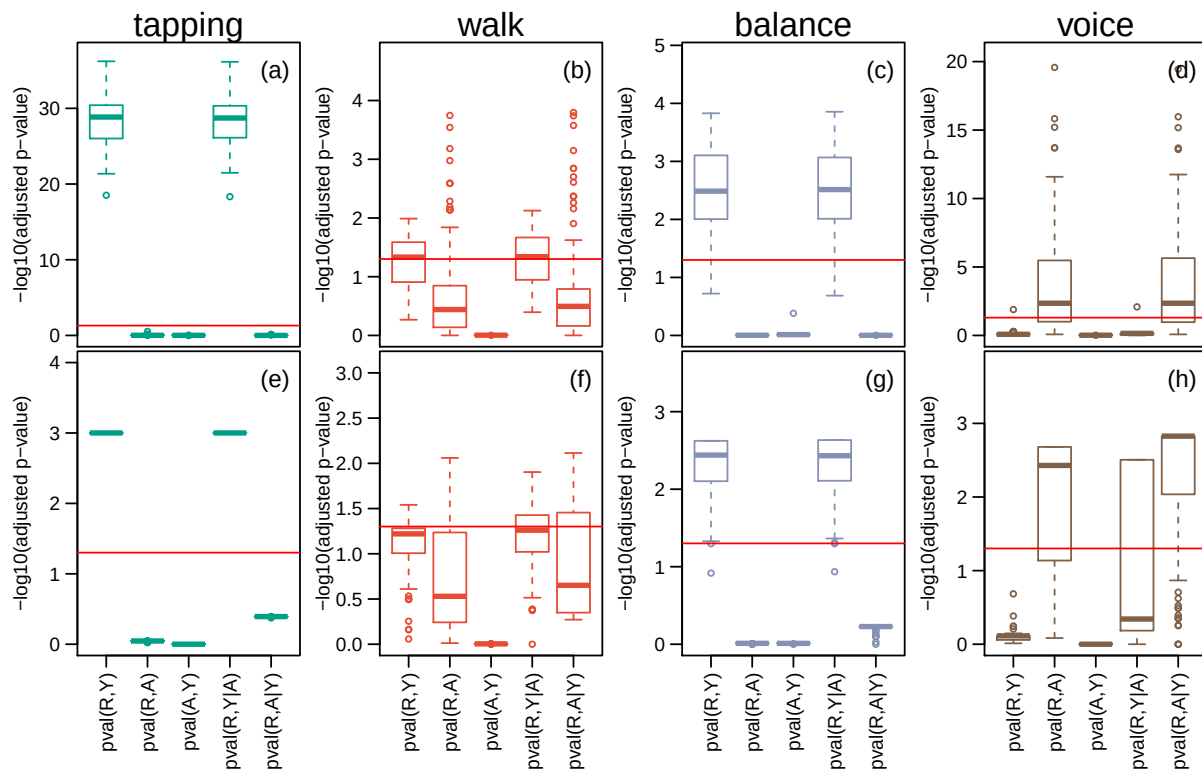
Supplementary Figure 27: Comparison of (partial) correlation and (partial) distance correlation tests for education confounding in the predictions of a random forest classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are consistent. Analyses based on $n = 960, 464, 466,$ and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

age – cor. vs distance cor. tests – ridge regression



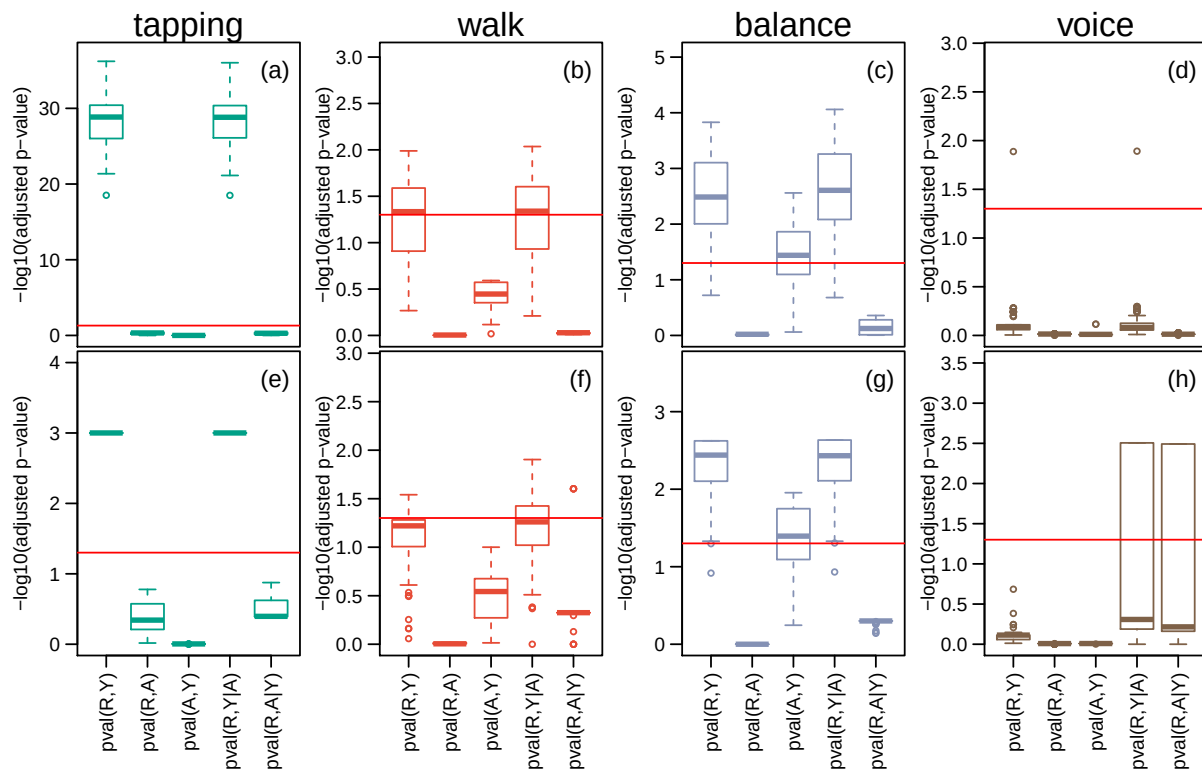
Supplementary Figure 28: Comparison of (partial) correlation and (partial) distance correlation tests for age confounding in the predictions of a ridge-regression classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are mostly consistent (except for a few data-splits on the voice assessment (panel h) which picked up a few non-linear associations missed by the correlations.) Analyses based on $n = 960, 464, 466$, and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

gender – cor. vs distance cor. tests – ridge regression



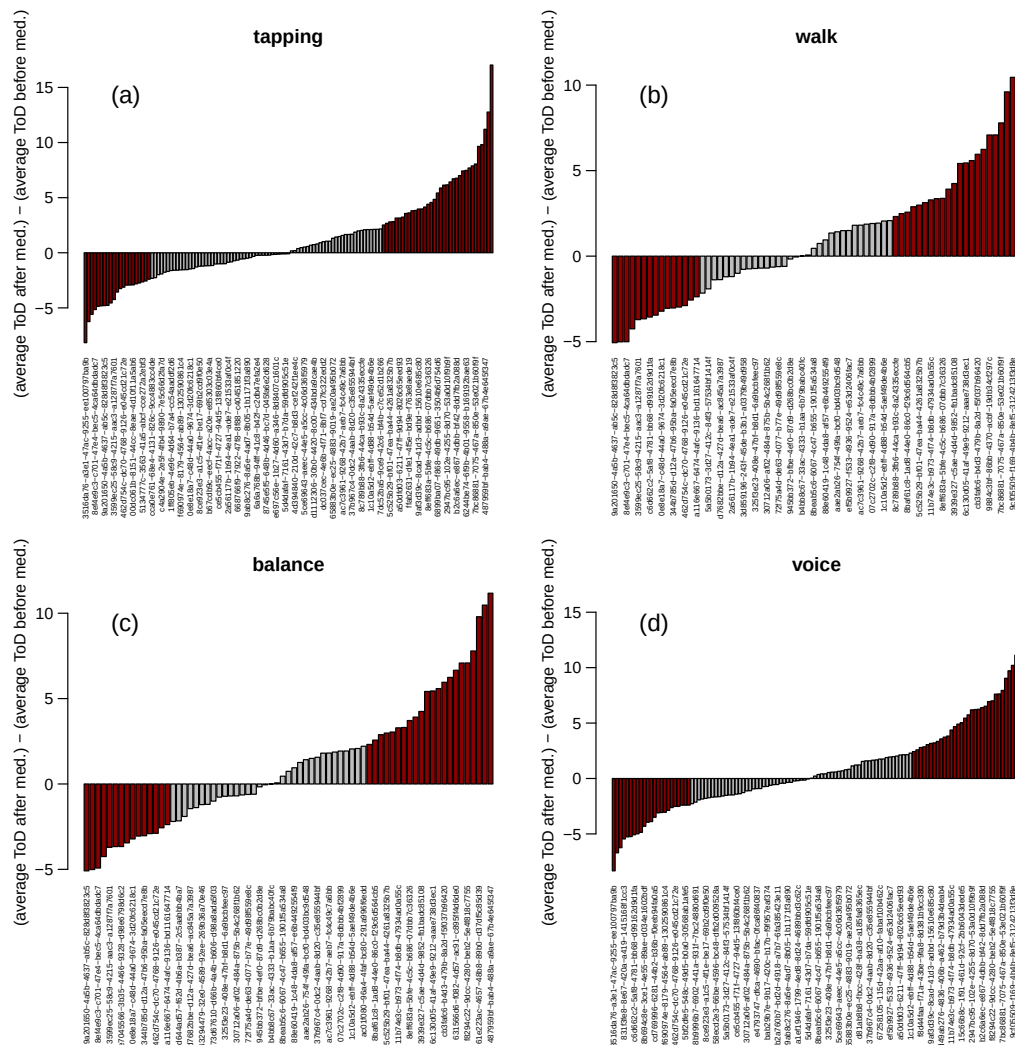
Supplementary Figure 29: Comparison of (partial) correlation and (partial) distance correlation tests for gender confounding in the predictions of a ridge-regression classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are mostly consistent (except for a few data-splits on the voice assessment (panel h) which picked up a few non-linear associations missed by the correlations.) Analyses based on $n = 960, 464, 466$, and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

education – cor. vs distance cor. tests – ridge regressio



Supplementary Figure 30: Comparison of (partial) correlation and (partial) distance correlation tests for education confounding in the predictions of a ridge-regression classifier trained with matched data. Top panels show the results from the two-tailed (partial) correlation tests. Bottom panels show the results from the one-tailed (partial) distance correlation tests (based on 1,000 permutations of the data). The results are mostly consistent (except for a few data-splits on the voice assessment (panel h) which picked up a few non-linear associations missed by the correlations.) Analyses based on $n = 960, 464, 466$, and 620 , for the tapping, walk, balance, and voice activities, respectively. Adjusted p-values were computed with Benjamini-Hochberg multiple testing correction. The data are represented as boxplots where the box corresponds to the the interquartile range (IQR), defined as the difference between the third quartile (Q3) and the first quartile (Q1), the middle line represents the median, the upper and lower whiskers are given by $Q3 + 1.5 \text{ IQR}$ and $Q1 - 1.5 \text{ IQR}$, respectively, and outlying points are plotted individually beyond the whiskers, with the lowest and highest points representing, respectively, the minimum and maximum.

3 Supplementary Figure 31



Supplementary Figure 31: Flagging individuals that show associations between time-of-the-day and before/after medication activities. In each panel, the x-axis presents the participants, while the y-axis shows the difference between the average time of all activities performed “after medication” versus the average time of all activities performed “before medication”. The results were sorted according to the value of the t-statistic used to test if the difference in times across the “before” and “after” medication groups were statistically significant. The red bars represent differences that were significant at a 0.05 level (after FDR multiple testing correction), while the grey bars show the non-significant differences. The red bars to the right of the grey bars represent individuals that tended to consistently perform the “before medication” activities earlier in the day than the “after medication” activities. The red bars to the left of the grey bars represent individuals that tended to perform the “before medication” activities later in the day than the “after medication” activities. Results were based on the data used for the personalized medication response analyses.

References

- [1] Chaibub Neto, E., et al. Causality-based tests to detect the influence of confounders on mobile health diagnostic applications: a comparison with restricted permutations. In Machine Learning for Health (ML4H) Workshop at NeurIPS 2019 - Extended Abstract. arXiv:1911.05139 (2019).
- [2] Pearl, J. *Causality: models, reasoning, and inference*. Cambridge University Press New York, NY (2009).
- [3] Pearl, J. and Mackenzie, D. *The book of why: the new science of cause and effect*. Basic Books, New York, NY (2018).
- [4] Spirtes, P., Glymour, C., & Scheines, R. *Causation, Prediction and Search*, 2nd ed. MIT Press, Cambridge, MA.
- [5] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300 (1995).
- [6] Szekely, G., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35**: 2769-2794 (2007).
- [7] Szekely, G., and Rizzo, M. L. Partial distance correlation with methods for dissimilarities. *Annals of Statistics* **42**: 2382-2412 (2014).