Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/11/eaaw9011/DC1

# Supplementary Materials for

## How to detect high-performing individuals and groups: Decision similarity predicts accuracy

R. H. J. M. Kurvers*, S. M. Herzog, R. Hertwig, J. Krause, M. Moussaid, G. Argenziano, I. Zalaudek, P. A. Carney, M. Wolf

*Corresponding author. Email: kurvers@mpib-berlin.mpg.de

Published 20 November 2019, *Sci. Adv.* **5**, eaaw9011 (2019)
DOI: 10.1126/sciadv.aaw9011

**The PDF file includes:**



**Other Supplementary Material for this manuscript includes the following:**

(available at advances.sciencemag.org/cgi/content/full/5/11/eaaw9011/DC1)

Skin cancer data set (.csv format).
R Code numerical simulations (Fig. 2B) (.R format).

## Supplementary text

### Structured correlations can break down the similarity-accuracy relationship: an example

When investigating the consequences of correlated decisions in our numerical simulations, for simplicity, we have assumed that these correlations do not differ systematically between subgroups of decision makers and/or cases. In the presence of such systematic differences, scenarios are conceivable where the positive correlation between decision similarity and accuracy is not observed even when the average accuracy of decision makers exceeds 0.5. To illustrate, consider the scenario where one individual always makes the correct decision, and all others perform imperfect, but above chance with accuracy $p$, $0.5 < p < 1.0$. Assume next that, for any given case, all imperfect individuals always make the same decision. In such a situation, the expected average decision similarity of the perfect decision maker is $p$ while that of the imperfect decision makers is an affine combination of $p$ and 1 ($\frac{1}{N-1} \cdot p + \frac{N-2}{N-1} \cdot 1$) which always exceeds $p$, even though the average accuracy in the population exceeds 0.5.

### Similarity to majority

As in the main text, we consider a pool of $N$ decision makers. For each individual $i$, $i = 1...N$, we define the decision similarity to the majority $S_i^m$ as the frequency of cases for which individual $i$ makes the same decision as the majority of the $N$ decision makers. The expected value of this similarity to the majority is given by

$$E(S_i^m) = p_i \cdot M\left|C_i + (1 - p_i) \cdot (1 - M\right|W_i) \quad \text{(S1)}$$

where $M\left|C_i\right.$ and $M\left|W_i\right.$ correspond to the probabilities with which the majority of the $N$ decision makers is correct, conditional on individual $i$ being correct or wrong, respectively. When comparing the decision similarity to the majority of two individuals $j$ and $i$, we find

$$E(S_j^m) - E(S_i^m) = (p_j - p_i) \cdot (2 \cdot M^- - 1) \quad \text{(S2)}$$

where $M^-$ is the probability with which – after excluding the individuals *j* and *i* from the pool of *N* individuals – the majority of the remaining pool of *N*-2 individuals is correct. As can be seen from Eq. (S2), and analogous to our result in the main text (Eq. (4) and (5)), whenever the majority of the remaining pool of individuals (i.e., excluding individuals *j* and *i*) is more often correct than incorrect (i.e., $M^- > 0.5$), the individual with the higher accuracy is characterized by a higher expected similarity to the majority, that is

$$p_j > p_i \Leftrightarrow E(S_j^m) > E(S_i^m) \qquad \text{(S3)}$$

**Accuracy of individuals with the highest similarity values vs. accuracy of the majority rule**

Selecting the individuals with the highest similarity values – be it similarity to others or similarity to the majority (see main text) – does not imply that these approaches aim to approximate the performance of the majority rule. To see this, note that similarity correlates with decision accuracy. Selecting the most similar individual thus aims at selecting the most accurate individuals from a pool of candidate decision makers. The most accurate individuals, in turn, often achieve a higher accuracy than the majority rule in relatively small groups with relatively large differences in accuracy (Example 1), while the majority rule often outperforms the most accurate individuals in small groups with small differences in accuracies (Example 2) and in large groups.

Example 1: Consider three decision makers with accuracies 0.55, 0.55 and 0.9. Here, the majority rule achieves an accuracy of 0.75 and the expected average decision similarities of the three decision makers to others are 0.52, 0.52 and 0.54, respectively (expected decision similarity to the majority: 0.75, 0.75 and 0.79, respectively). Thus, in this scenario, the individual with the highest similarity (i.e., the most accurate decision maker) outperforms the majority rule.

Example 2: Consider three decision makers with accuracies 0.7, 0.75 and 0.8. Here, the majority rule achieves an accuracy of 0.85 and the expected average decision similarities of the three decision makers to others are 0.61, 0.63 and 0.64, respectively (expected decision similarity to the majority: 0.79, 0.82 and 0.84, respectively). Thus, in this scenario, the individual with the highest similarity (i.e., the most accurate decision maker) performs worse than the majority rule.

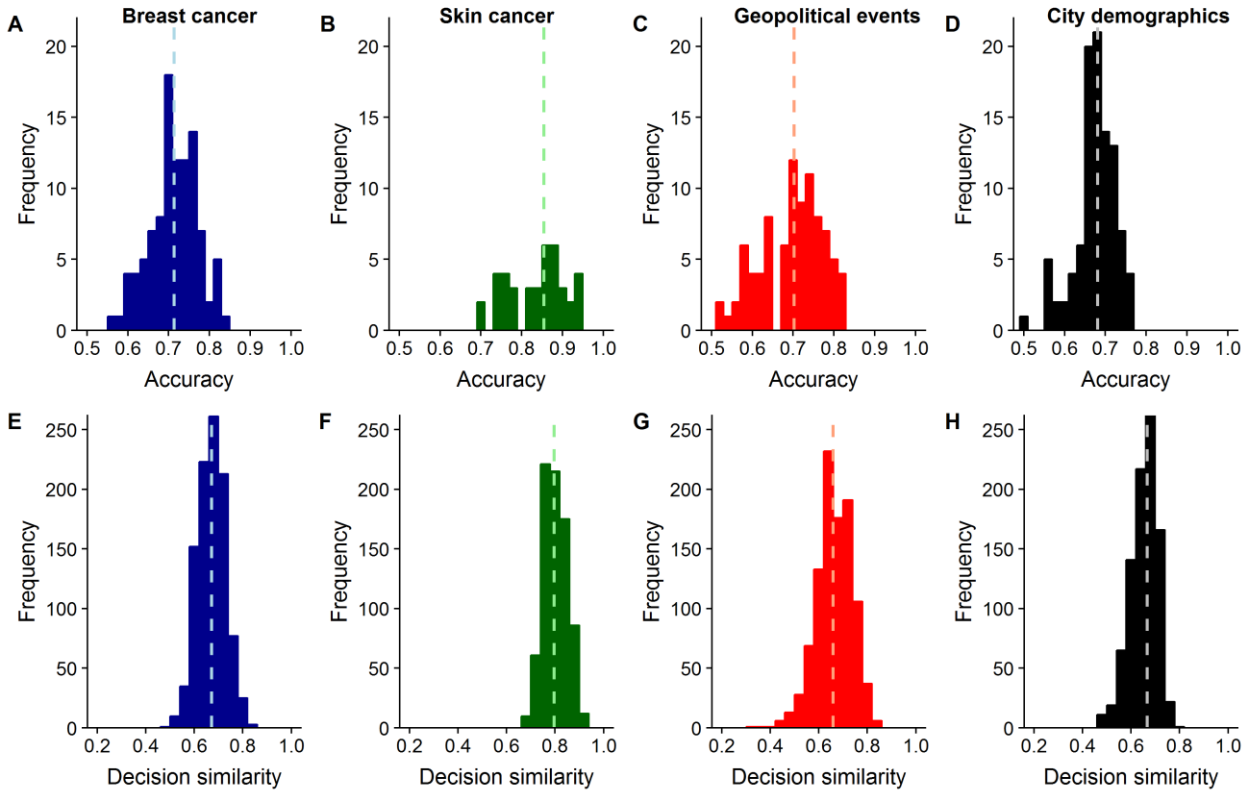


**Fig. S1. Distribution of individuals' level of accuracy and correlated decisions in the four datasets.** (**A-D**) Frequency distribution of the accuracy of individuals in each data set. In (A, B) accuracy is expressed as balanced accuracy; and in (C, D) as proportion correct. (**E-H**) Frequency distribution of decision similarity (i.e., percentage agreement) between randomly drawn pairs of individuals within each data set, sampling a maximum of 1,000 unique pairs. Across the four data sets, individuals showed, on average, fair to moderate agreement. Dashed vertical lines depict median values.

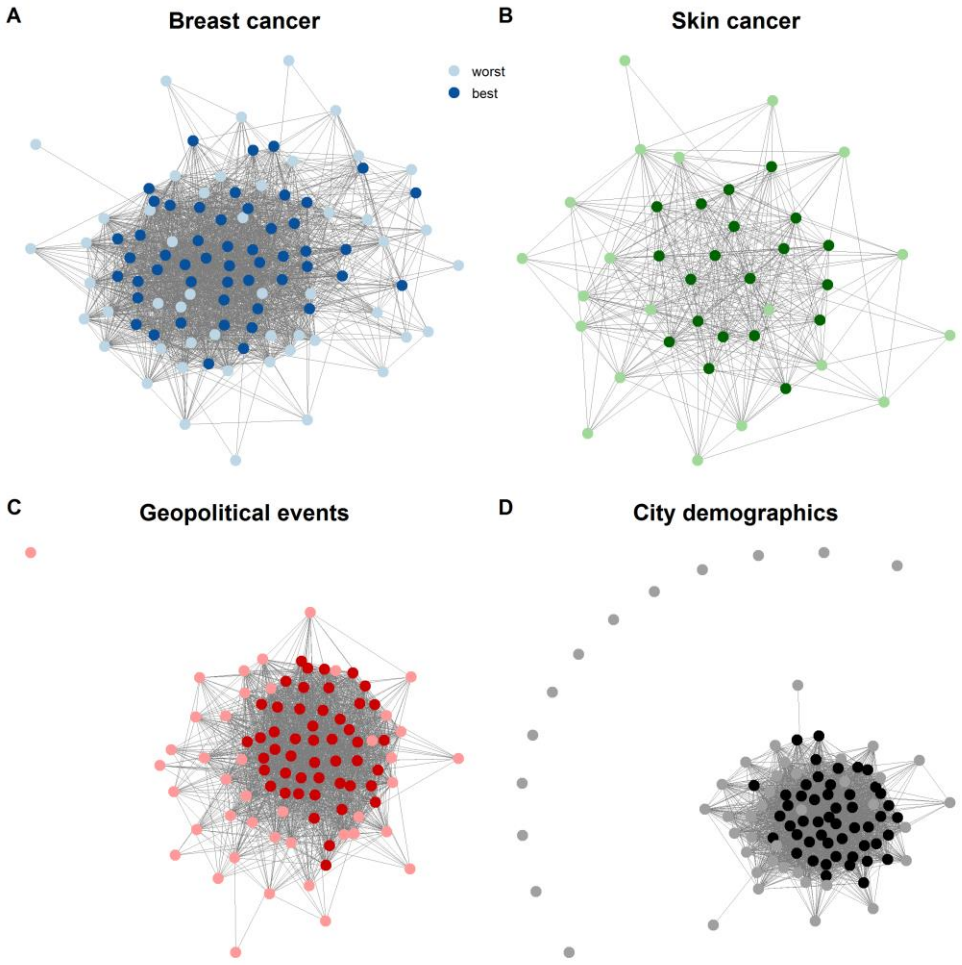


**Fig. S2. High-performing individuals are similar to each other, while low-performing individuals tend to make dissimilar decisions.** (**A-D**) To further study the relationship between individuals' accuracy levels and decision similarity with others on a population level, we created, for each data set, a network in which nodes represent individuals (nodes are colored according to individual accuracy: dark/light shades = 50% most/least accurate individuals) and edges represent above median decision similarity between raters. That is, lines connect raters that have a relatively high level of agreement in their decisions. We used the Kamada-Kawai algorithm, which minimizes the distance between pairs of individuals with high decision similarity. Decision similarity between individuals was thus the sole basis for creating the network. As can be seen, in all data sets, high-performing individuals cluster in the center of the network, whereas low-performing individuals are scattered at the network periphery. Low performers thus not only made dissimilar decisions to high performers but also to other low performers; that is, while high performance unfolds in a similar way, poor performance unfolds in different ways. (C, D) Several nodes are isolated from all other nodes, implying that the pairwise similarity values of these individuals with any other individual in the population are below the median.

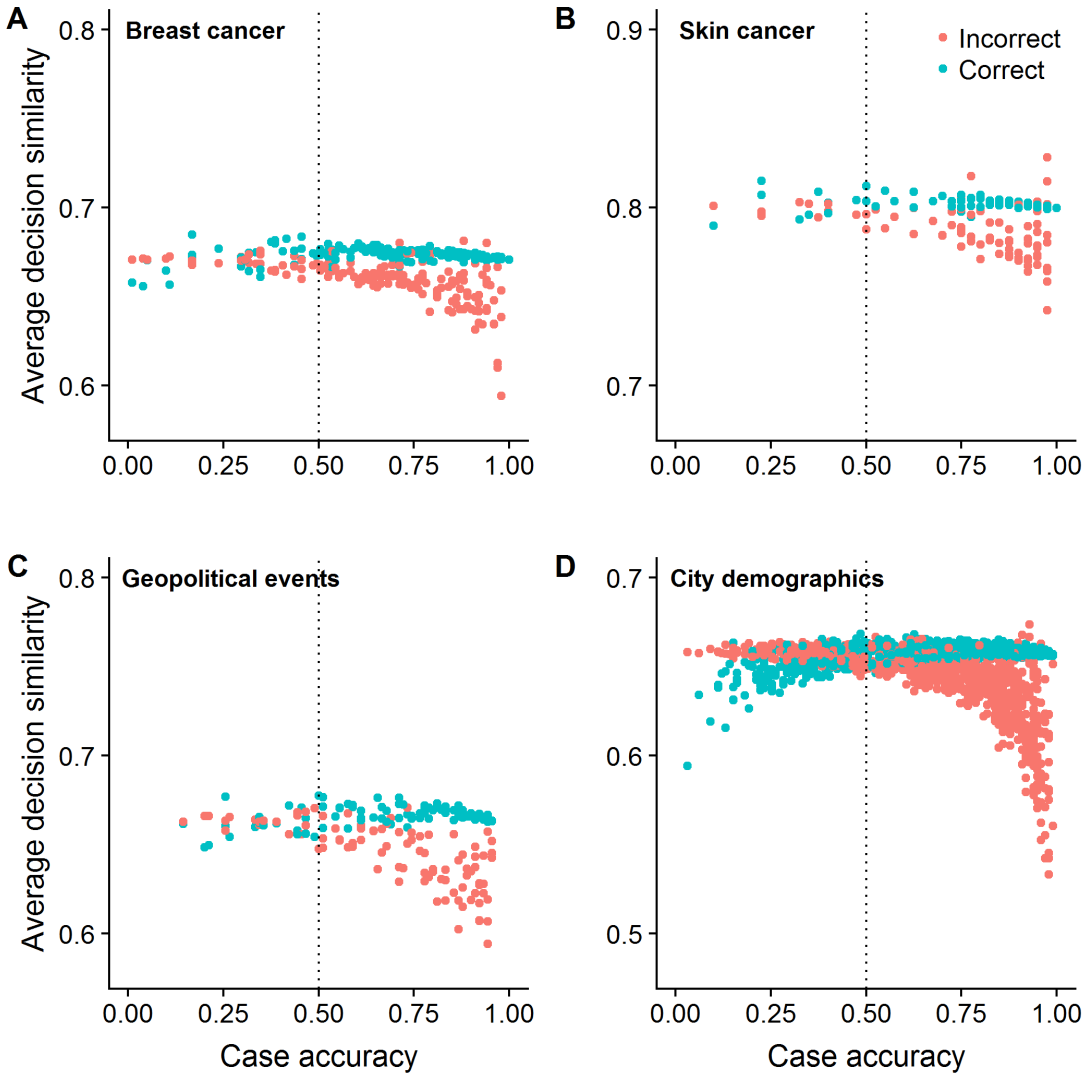


**Fig. S3. Decision similarity performs well for cases in which the majority decided correctly but breaks down for cases in which the minority decided correctly.** (**A-D**) The relationship between case accuracy (i.e., the fraction of the population answering a particular case correct) and average decision similarity for individuals who were correct (blue dots) and who were incorrect (red dots) on that particular case. Each case is thus represented by two dots, one for individuals being correct, and one for individuals being incorrect on that case. For cases in which the majority was correct (case accuracy > 0.5), the individuals being correct had higher similarity values than the individuals being incorrect; for cases in which the minority was correct (case accuracy < 0.5), this pattern either disappeared (A-C), or even reversed (D).

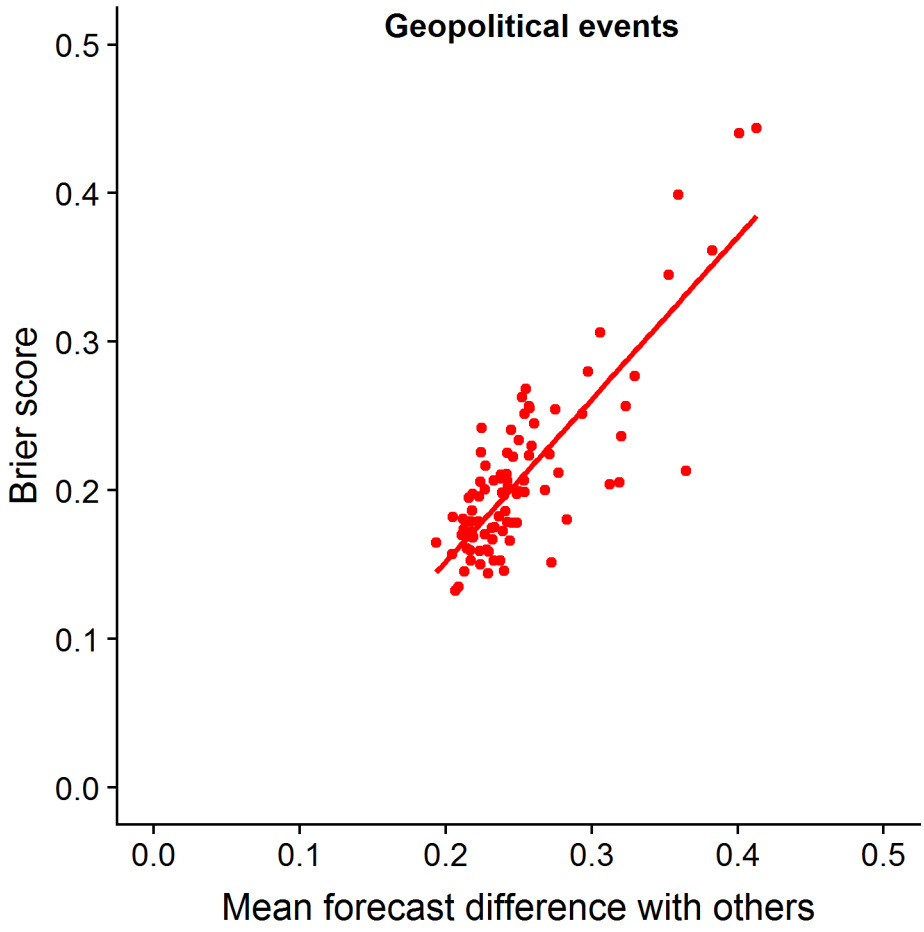


**Fig. S4. The similarity-accuracy relationship is also present when using the continuous probability forecasts.** To investigate the relationship between similarity and individual accuracy of continuous probabilistic forecasts (rather than binary yes/no forecasts), for each individual, we calculated its mean absolute distance between its own probabilistic forecasts and that of all others. Next, we calculated the average accuracy of each individual using the Brier score, where lower values indicate higher accuracy. As can be seen, analogous to our results for binary forecasts (Fig. 3C, main text), the closer an individual's forecasts were to other forecasts, the higher its accuracy (i.e., the lower its Brier scores).

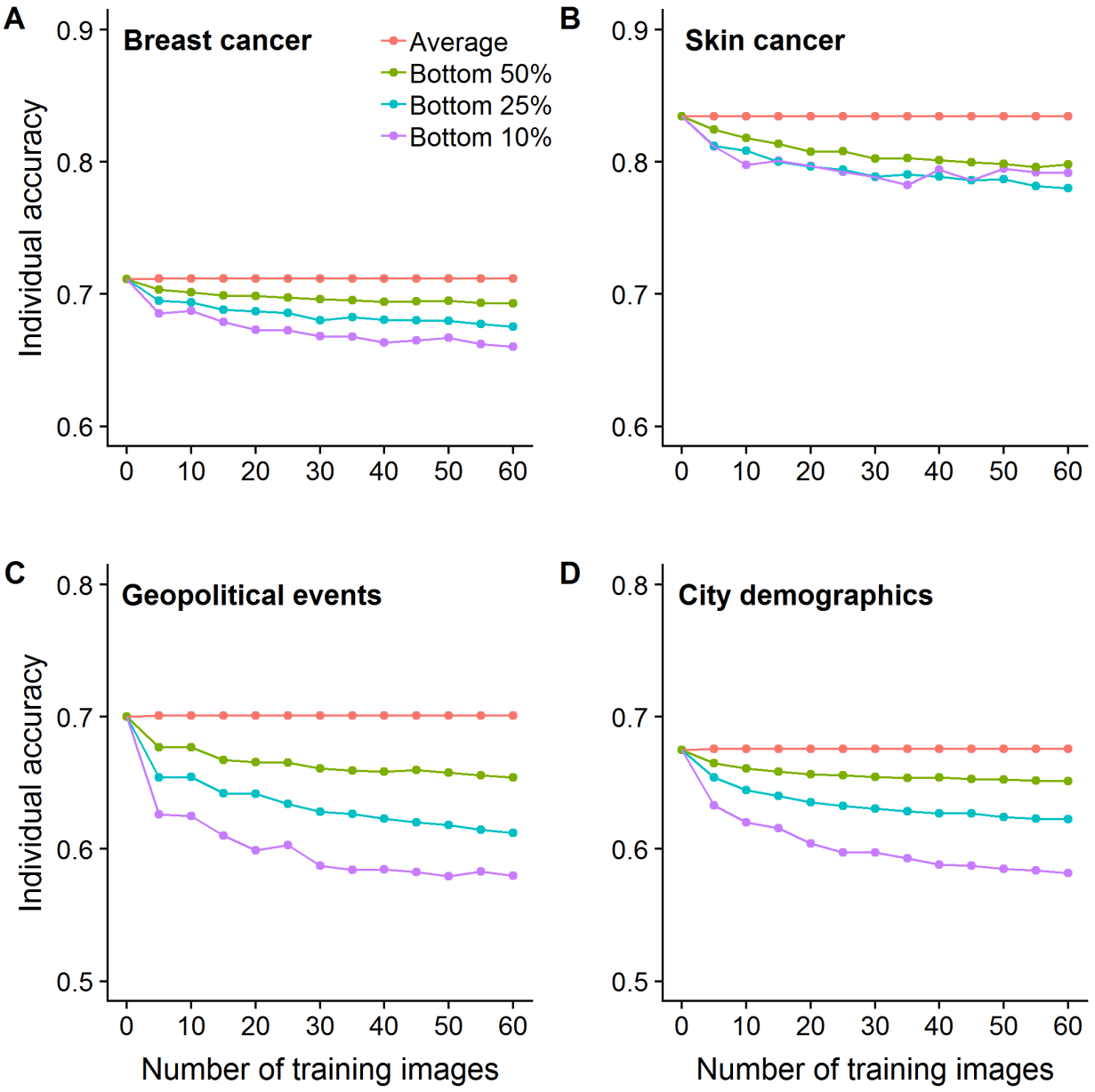


**Fig. S5. Decision similarity permits identification of low-performing individuals.** (**A-D**) The average performance in a test set of individuals selected on the basis of their decision similarity in a training set, for different decision similarity thresholds (e.g., the bottom 25% corresponds to the 25% of raters with the lowest decision similarity in the training set) and different numbers of training images. The performance of the selected individuals in the test set decreases with a decreasing threshold value and with increasing the number of training images, reflecting an increase in the correct identification of low-performing individuals. The red line ("Average") refers to the average performance of all individuals in the test set. In (A, B) accuracy is expressed as balanced accuracy; and in (C, D) as proportion correct.

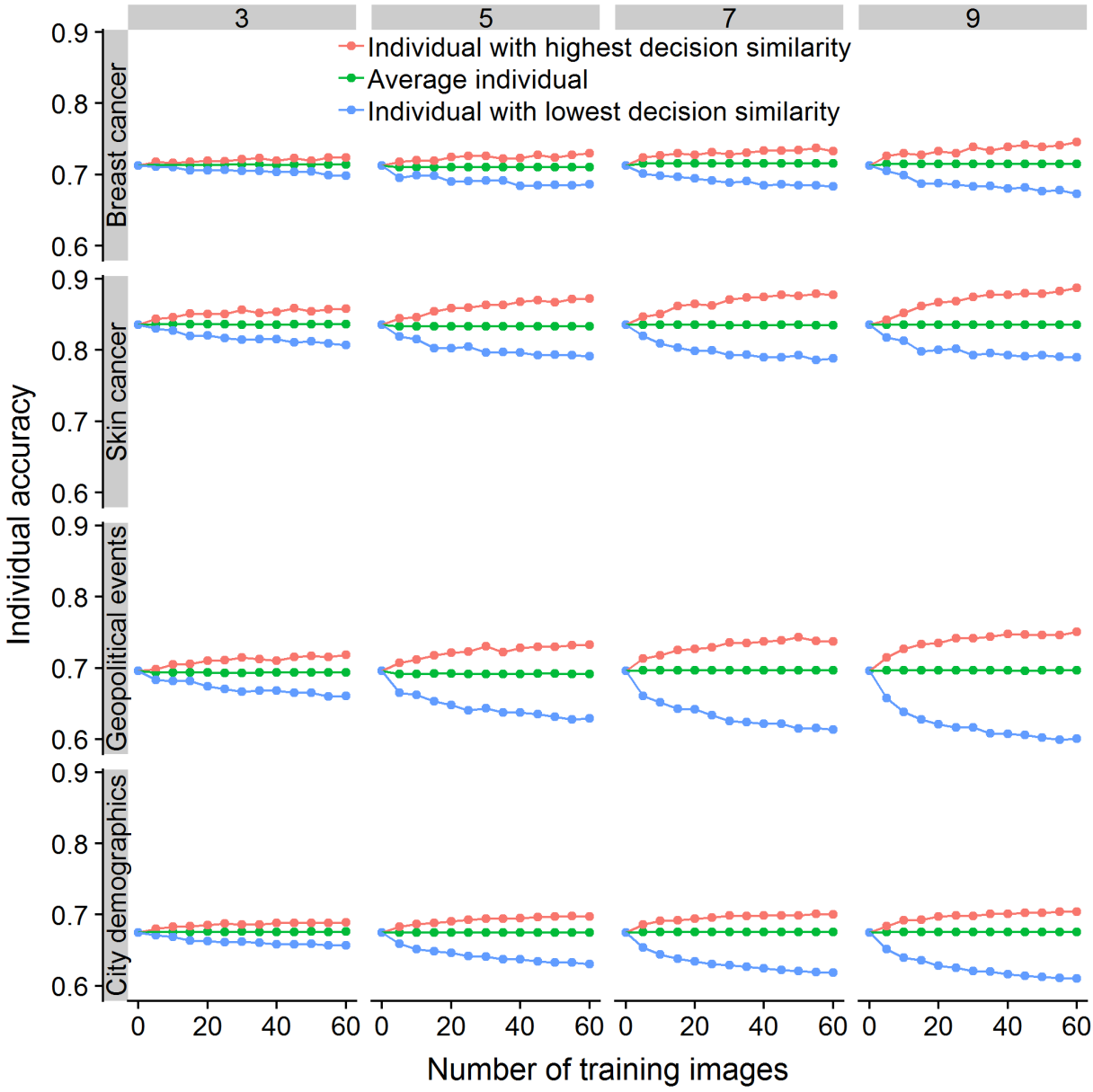


**Fig. S6. Decision similarity permits identification of high-performing (and low-performing) individuals in small groups.** To test whether decision similarity can be used to identify high- (and low-) performers in small groups, we performed cross-validation procedures using small groups of different sizes. Within each data set, we randomly sampled *n* individuals (3, 5, 7, and 9). We then randomly sampled *m* cases to create a training set (varying *m* from 0 to 60 in 5-step increments), and used part of the remaining cases to form a test set (see Materials and Methods for details). In the training set, we calculated, for each individual, its average decision similarity. We then selected the individual with the highest and the one with the lowest decision similarity and tested the performance of both in the test set. We replicated each combination of group size and number of training images 1,000 times in each data set. The green line ("Average") refers to the average performance of all individuals in the test set.

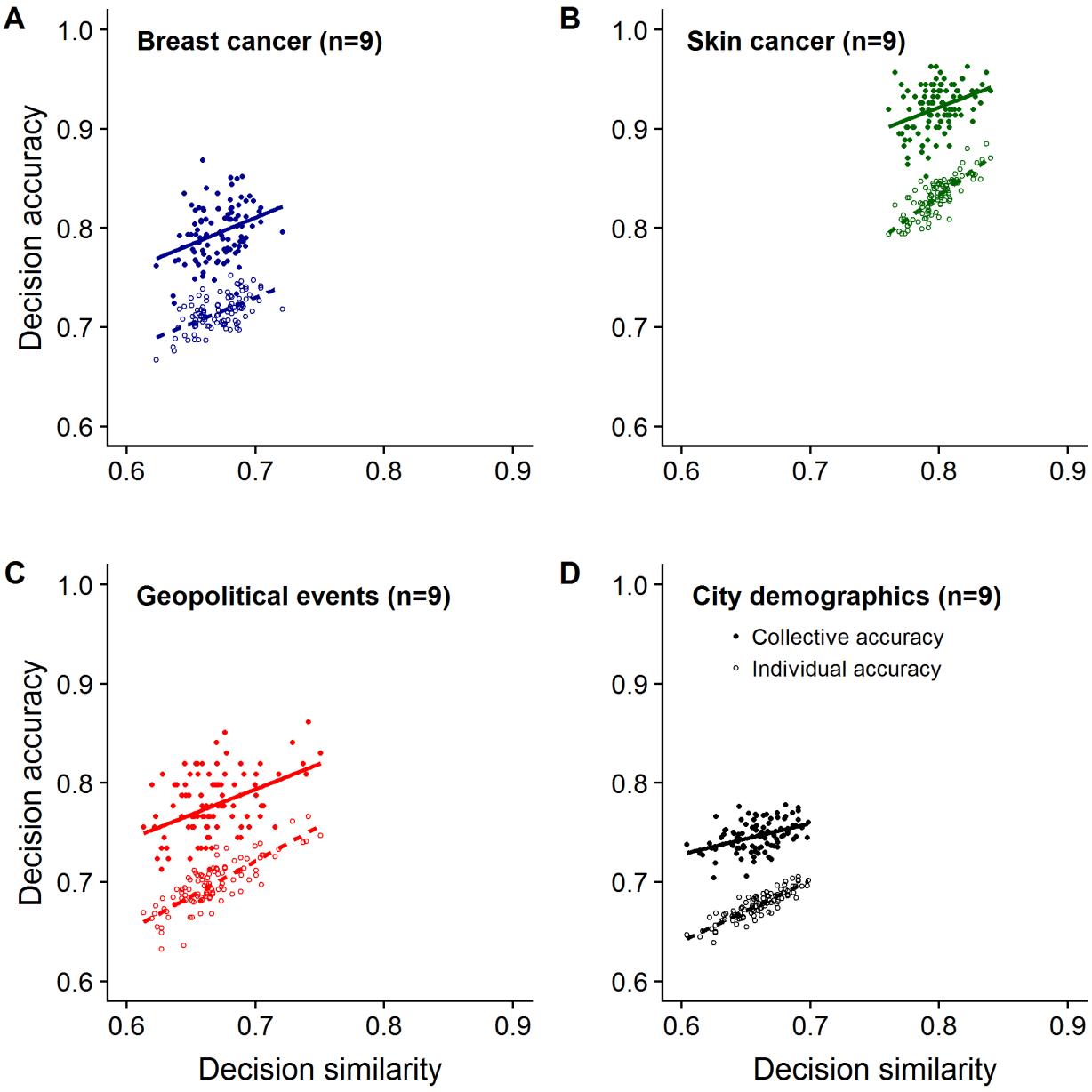


**Fig. S7. The relationship between decision similarity of a group of nine individuals and their individual and collective accuracy.** (**A-D)** In each of the data sets, we find a strong positive relationship between the average decision similarity among group members and their collective performance under the majority rule (filled dots and solid regression lines). This pattern is driven by a strong positive relationship between the average decision similarity among group members and the average individual performance of group members (open dots, dashed regression lines). In (A, B) accuracy is expressed as balanced accuracy; and in (C, D) as proportion correct.

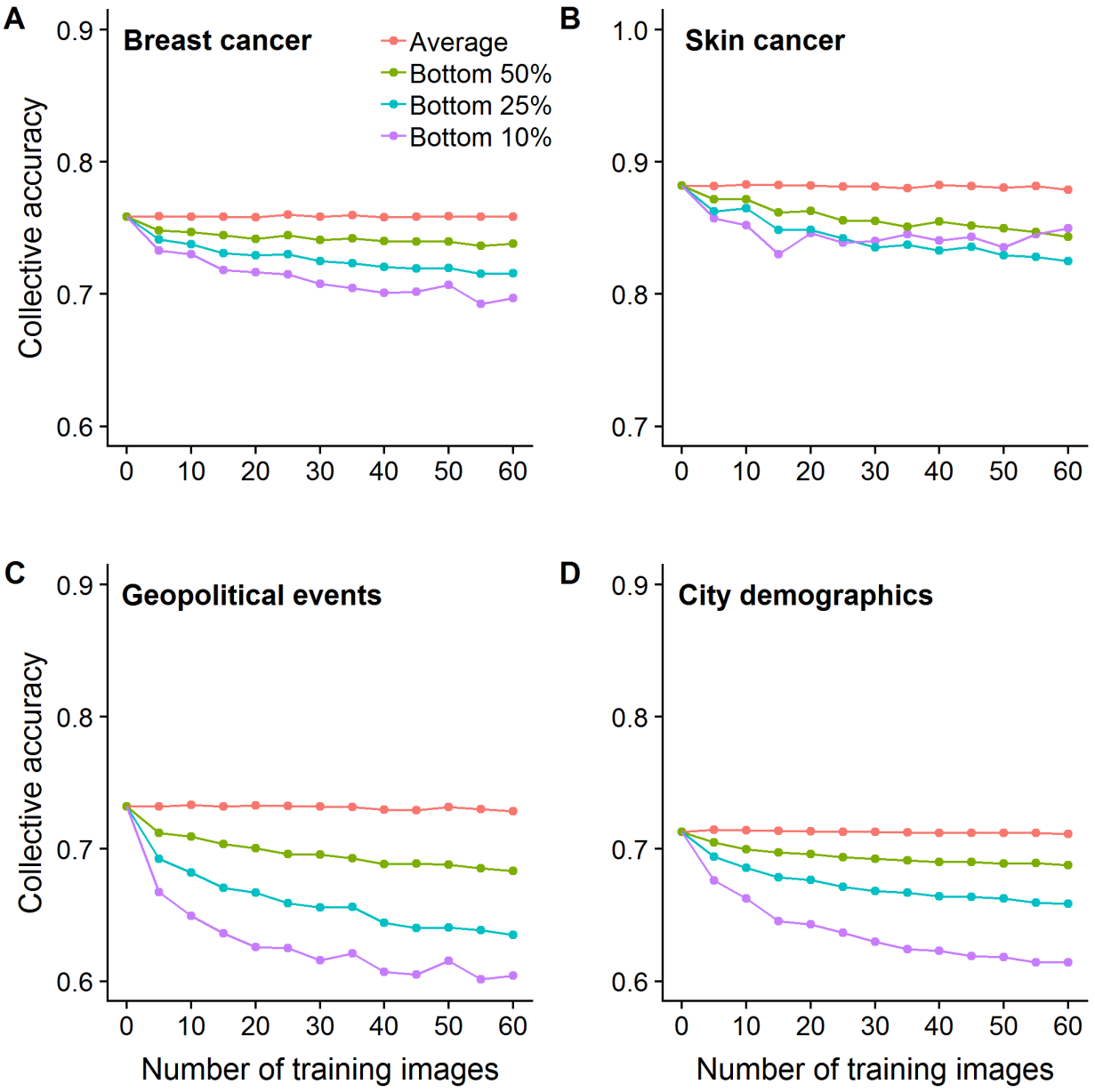


**Fig. S8. Decision similarity permits identification of low-performing groups.** (**A-D**) The collective performance in a test set, using a majority rule, of groups of individuals selected on the basis of the decision similarity in a training set, for different similarity thresholds (e.g., the bottom 25% corresponds to groups containing individuals with the 25% lowest decision similarity values) and different numbers of training images. The performance of the selected groups of individuals in the test set decreases with decreasing the threshold value and increasing the number of training images, reflecting an increase in the correct identification of low-performing groups. The red line ("Average") refers to the average majority-rule performance of all groups in the test set. In (A, B) accuracy is expressed as balanced accuracy; and in (C, D) as proportion correct.

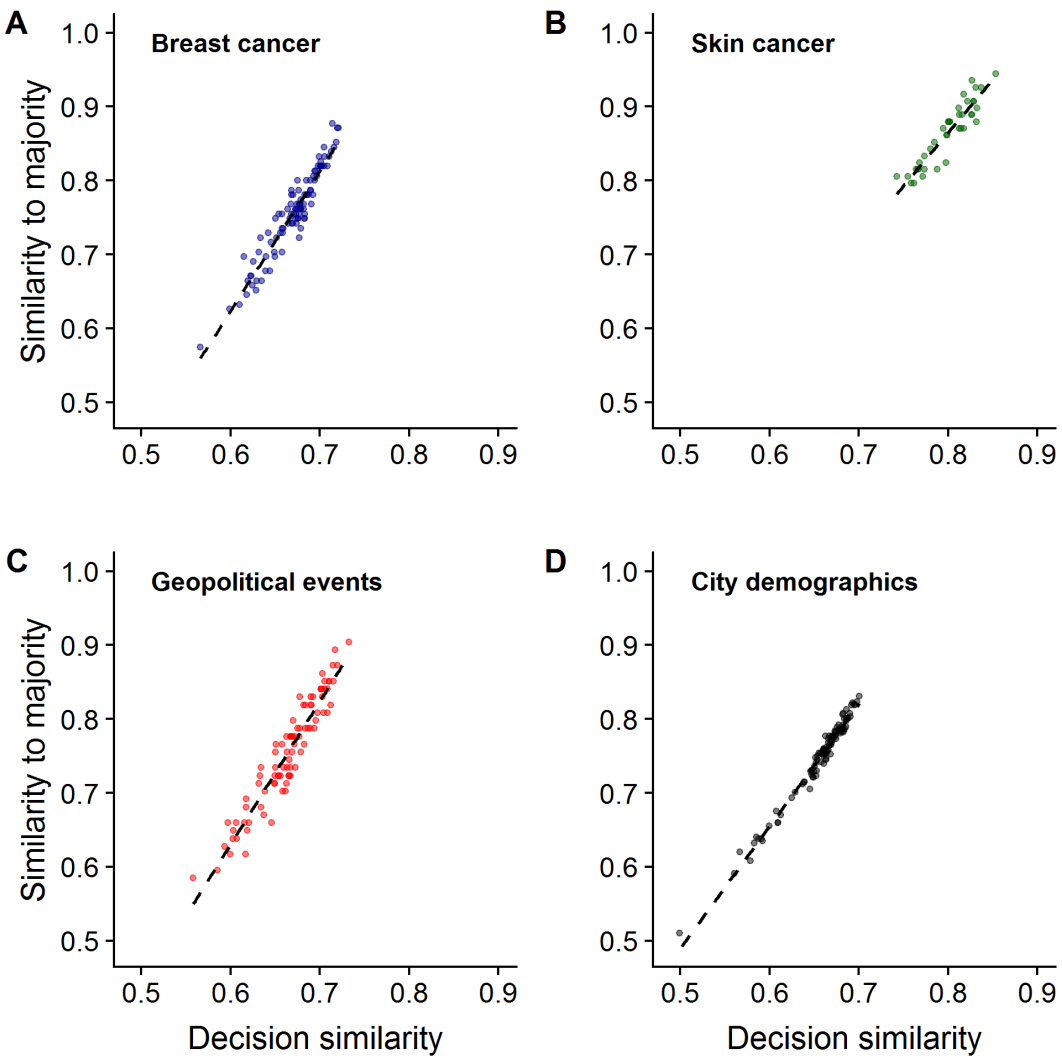


**Fig. S9. In each of the four datasets, the average decision similarity to others tightly correlates with the decision similarity to the majority judgment.** Each dot corresponds to one individual. To determine how an individual's decision similarity to others relates to that individual's decision similarity to the majority, we calculated, within each data set, for each individual, its decision similarity to the majority rule (i.e., the percentage agreement with the majority outcome). Values of 0.5 and 0.7, for example, imply that the individual makes the same decision as the majority in 50% and 70% of the cases, respectively. As can be seen, individuals with a high average decision similarity to others also have a high decision similarity to the majority judgment.

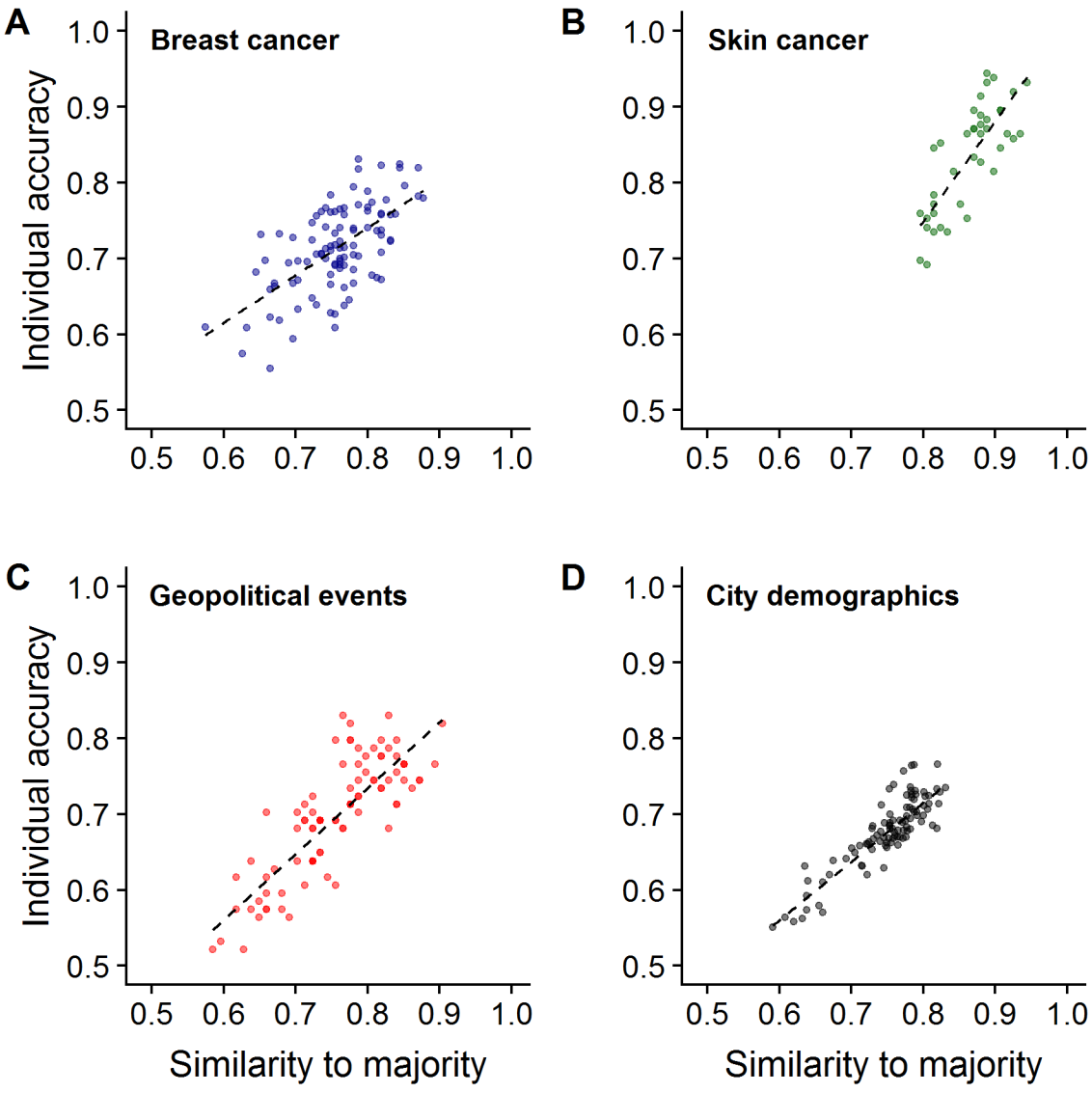


**Fig. S10. Decision similarity to the majority tightly correlates with decision accuracy in breast and skin cancer diagnostics, geopolitical forecasting, and a general knowledge task.** (**A-D**) In all four data sets, we find a positive relationship between individuals' decision similarity to the majority (i.e., average percentage of agreement with the majority outcome) and accuracy. In (A, B) accuracy is expressed as balanced accuracy and in (C, D) as proportion correct. Lines are robust linear regression lines.