Towards more reliable and fairer decision-making systems: pooling decisions decreases variation in accuracy and response bias

Ralf H. J. M. Kurvers[1,2]*, Stefan M. Herzog[1], Ralph Hertwig[1], Jens Krause[2] & Max Wolf[2]

[1] Center for Adaptive Rationality, Max Planck Institute for Human Development. Lentzeallee 94, 14195 Berlin, Germany.

[2] Leibniz Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310, 12587 Berlin, Germany.

* Corresponding author:

Email: kurvers@mpib-berlin.mpg.de

Keywords: Collective decision making, groups, medical diagnostics, forecasting, fairness

Abstract

Over the last decades, the relative benefits and costs of individual vs. collective decision-making systems have attracted ample attention in the behavioural sciences and beyond. This research however, has almost exclusively focused on accuracy as a performance criterion, neglecting another major performance dimension of decision-making systems, the variation in outcomes between decision-making agents. This is surprising as low outcome variation is a key goal in many high-stake contexts, including medical, judicial and political decision making. Employing a combined theoretical and real-world data-driven approach, we investigate how one of the most prominent systems of collective decision-making – the pooling of independent decisions using the majority rule – affects the variation in outcomes between agents. Using a general statistical argument and large-scale numerical simulations, we predict that pooling decisions robustly reduces variation in two key outcome variables: accuracy and response bias (i.e. the decision maker's tendency towards one response or the other). We test this prediction in real-world datasets on breast and skin cancer diagnostics, fingerprint analysis, geopolitical forecasting, and a general knowledge task, encompassing more than 350 decision makers making more than 125,000 decisions. As predicted, we find that pooling decisions robustly reduces variation in accuracy and response bias. Importantly, this reduction is accompanied by an increase in accuracy, showing that pooling independent decisions can simultaneously decrease variation and increase accuracy. Thus, while outcomes in individual decision-making systems are highly variable and at the mercy of individual decision makers, pooling decisions decreases this variation, thereby promoting more predictable, reliable and fairer outcomes.

Introduction

The process of making decisions is key to human societies, including decision making in the medical judicial, economic and political domain. The design of this process and the associated decision-making systems is thus of paramount importance to human welfare. One of the key properties of decision-making systems is whether the agent making the decision is a single or a collective of individuals. As a consequence, across the behavioural sciences, a large body of research has in recent decades focused on mapping out the relative advantages and disadvantages of individual vs. collective decision-making systems (1-5). Notwithstanding the substantial progress that has been made, the focus of this research has, almost exclusively, been on accuracy as the performance criterion (2, 3, 6-13) (but see 5), neglecting a second major performance dimension of decision-making systems, the variation in outcomes between decision-making agents. We here provide a first step to fill this important knowledge gap.

Individual decision makers differ in key aspects of the outcomes they produce and these differences matter. In many domains (e.g. medical diagnostics, truth detection, and geopolitical forecasting), it has been found that experts differ substantially in their level of accuracy (9, 14-18). Moreover, in binary classification tasks, like truth detection and many instances of medical diagnostics, next to accuracy, experts also differ in how much evidence they require to classify a case as either signal (e.g. cancer, lie) or noise (e.g. non-cancer, truth), a.k.a. response bias (19-22). Generally, the higher the variation in outcomes (i.e. accuracy levels, response bias) between decision makers, the less predictable and reliable the associated decision-making system is. Arguably, systems with low variation are perceived as fairer and more trustworthy than those where the outcome depends heavily on the specific decision maker. To see the importance of this, consider the equality-before-the-law principle. While this fundamental legal principle is typically meant to imply that judicial outcomes should not depend on irrelevant characteristics (e.g. race, or gender) of the person on trial, it also implies that outcomes should be independent of the specific agent (e.g. court, or jury) that prosecutes and decides on a defendant's case. Similarly, in the medical domain, patients facing a health threat hope that the diagnosis and treatment they receive is not subject to the idiosyncrasies of the physician they happen to encounter. Thus, in many contexts, like judicial and medical but also political decision-making, next to high decision accuracy, a low degree of variation in outcomes is an important benchmark for decision-making systems.

In the following, we use a combined theoretical and real-world data-driven approach to investigate how a simple, yet highly effective system of collective decision-making – the pooling of independent decisions using the majority rule – affects the variation in outcomes between decision-making agents in binary classification tasks like, for example, classifying a mammogram as “cancer present” or “cancer absent”. A first intuition can be derived from the law of large numbers: as group size becomes sufficiently large, the mean characteristics of the members of a group (e.g. the mean accuracy level or the mean response bias of group members) tend to approach the population mean, reducing variation between different large groups in this sense. We highlight, however, that we are here interested in variation in decision outcomes between (potentially small) groups and that, generally, decision outcomes from a group pooling independent decisions cannot be predicted from the mean characteristics of its members alone. We thus proceed in two steps. First, based on a statistical argument and large-scale numerical computer simulations, we derive general predictions about how pooling decisions affects variation in decision outcomes between decision-making agents in binary classification tasks. Second, we use several real-world datasets on breast and skin cancer diagnostics, fingerprint analysis, geopolitical forecasting, and a general knowledge task to empirically test these predictions.

Results

Statistical argument and numerical simulations

We start by developing a basic formal intuition about the relationship between pooling decisions and outcome variation in binary classification tasks. We begin with the most basic scenario where individual decision makers do not differ in accuracy levels or response bias. The task at hand is to solve a set of binary choice problems, for example, to classify a mammogram as “cancer present” or “cancer absent”, to decide whether a defendant is guilty or not, or to predict whether an earthquake will occur in a given region or not. We consider a population of individual decision makers, with each individual being characterized by the same accuracy level *a,* corresponding to the individual’s probability of being correct in any given choice. We further assume that the decisions of different individuals are statistically independent from each other.

Although individuals have identical accuracies, they may still differ (stochastically) in how they respond to specific cases. The expected frequency of disagreement between two individuals can be calculated from the binomial distribution as $2 \cdot a \cdot (1-a)$. Focusing on situations where individuals achieve an accuracy level above chance (i.e. $a > 0.5$), we make two observations. First, the higher the accuracy *a* of two individuals, the lower their expected frequency of disagreement (the derivative of the expected disagreement with respect to *a* is $2 - 4 \cdot a$, which is negative as long as $a > 0.5$). Second, groups pooling the independent decisions with a majority rule achieve a higher accuracy than individuals, with larger groups achieving a higher accuracy than smaller groups (Condorcet Jury Theorem) (23). From these two observations follows that – in this most basic scenario – pooling decisions decreases outcome variation between agents: as group size increases, accuracy increases thereby decreasing the expected frequency of disagreement between groups (an analogous logic applies to situations with $a < 0.5$, where pooling decreases accuracy and thereby decreases disagreement). Figure 1 illustrates this relationship for different individual accuracy levels and different group sizes. As can be seen, for any given accuracy level *a* – with the exception of accuracies 0.0, 0.5 and 1.0 – the variation between agents (i) is larger between individual decision makers (red line) than between groups of decision makers; (ii) decreases as group size increases; and (iii) approaches 0 as group size becomes sufficiently large, with the exception of accuracy 0.5. This general statistical argument, concerning the most basic case (i.e. absence of individual differences in accuracy and response bias), suggests that pooling decisions could be a powerful approach to reduce variation in outcomes between decision-making agents.

We next investigate the consequences of pooling decisions for variation in more realistic scenarios, with decision makers differing in accuracy levels and response bias. Importantly, whenever individuals differ in accuracy levels and/or response bias, the decision outcomes produced by a group employing the majority rule (and thus the variation in outcomes between groups) will depend on the specific composition of individuals in that group. We thus use numerical simulations to investigate these more complex scenarios.

We first focus on differences in accuracy levels. In order to do so, we consider a large range of different populations of decision makers that differ in their accuracy distribution by systematically varying the mean and the variance of these beta distributions (Fig. 2A). As a result, these populations differ in their average accuracy, variance, and skewness. For each of these populations, we repeatedly and randomly

sample groups of *n* decision makers (i.e. individuals characterized by a given accuracy level, varying *n* from 1, 3, 5, to 9). Each of these *n* decision makers within one group faces the same *m* decision cases (varying *m* from 10, 25, to 100), and they independently rate the *m* cases, based on their accuracy level. Next, we pool the decisions of the *n* decision makers for each of the *m* cases, using a majority rule: If the majority of decision makers rates the case correctly (incorrectly), the collective decision is classified as correct (incorrect). We then calculate the group performance by summarizing the majority outcomes across the *m* cases (i.e. proportion correct decisions). For each unique combination of group size *n*, cases *m*, and accuracy distribution we sample 10,000 groups. Finally, for each unique treatment combination, we calculate the variance in group accuracy among the corresponding 10,000 groups as a measure of outcome variation between decision-making agents.

Figure 2B shows the results of this analysis. Within each subpanel, the different tiles correspond to the different populations (i.e. accuracy distributions) of the associated tiles in Fig. 2A. Within each subpanel, moving from left to right corresponds to an increase in the mean accuracy of populations; moving from bottom to top corresponds to an increase in the variance in individual accuracy. Within each subpanel, the middle column (i.e. mean accuracy = 0.5) corresponds to rater populations performing at chance level, and tiles to the left (right) indicate populations performing below (above) chance. Importantly, as can be seen from Fig. 2B, independent of the specific accuracy distribution, we find that the variation in accuracy between individual decision makers is larger than the variation between groups of decision makers and that increasing group size reduces the variation in performance between groups (i.e. different decision-making agents). This effect is smallest for populations at chance level, and increases the closer mean accuracy moves towards 1 or 0. Importantly, the reduction in variation in accuracy between groups with increasing group size is present in almost all scenarios. This strongly suggests that the reduction in outcome variation in accuracy with larger groups is a robust effect. *SI Appendix, Fig. S1* shows that, as expected, the reduction of variation in accuracy between groups increases even further with increasingly larger groups.

We next focus on the consequences of pooling decisions for variation in response bias. In order to do so, we consider scenarios where the world can be in two states (e.g. cancer present or absent; defendant guilty or not, etc.) and decision makers can thus make two types of errors (22, 24-26). In order to reduce the number of possible scenarios to a feasible set, we make two simplifying assumptions: first, both states

of the world appear equally often; second, decision makers within a population have the same ability to discriminate between positive and negative cases. Following the simplest signal detection approach for such a setup (equal-variance Gaussian model) *(22, 27)*, we use *d'* as a measure of discrimination ability (which we varied from 1 to 1.5 to 2) and the criterion value *c* as a measure of response bias (see *Materials and Methods*). Analogous to the accuracy analysis above, we generate a large range of different populations differing in their distribution of response biases, by systematically varying the mean and the variance of the associated criterion-distribution, allowing the criterion of individuals to range between –1 and +1 (Fig. 2C). For each of these populations, we repeatedly and randomly sample groups of *n* decision makers (i.e. individuals characterized by a particular criterion value, varying *n* from 1, 3, 5, and 9), fixing the number of cases to be decided upon to 1,000. For each simulated group pooling decisions with a majority rule, we calculate the accuracy (i.e., proportion correct) separately for state 1 (truly positive cases: sensitivity) and state 2 (truly negative cases: specificity) and use these values to calculate the implied criterion value (see *Materials and Methods*). For each combination of group size *n*, discrimination ability *d'* and criterion-distribution, we again independently sample 10,000 groups and calculate the variance in criterion among the corresponding 10,000 groups as a measure of variation in response bias.

Figure 2D shows the results of this analysis with darker colors indicating higher variance in response bias between groups. Within each subpanel, the different tiles correspond to the different populations of the associated tiles in Fig. 2C. Importantly, independent of the specific criterion-distribution and discrimination ability, increasing group size reduces the variation in response bias between decision-making agents. That is, based on a wide variety of populations with individual decision makers differing in response biases, we find that the pooling of individual decision makers' decisions substantially reduces variation in response bias between agents. This reduction is observed in almost all scenarios, thus suggesting this to be a robust effect. *SI Appendix, Fig. S2* shows that, as expected, increasing group size also leads to a reduction in variance in sensitivity and specificity between groups.

Empirical analysis

Our theoretical and simulation results suggest that a decision-making system that pools decisions is a powerful approach to reduce outcome variation between decision-making agents in binary classification

tasks. To test this prediction in real-world contexts, we next analyze several published datasets from five domains: (i) a breast cancer dataset, comprising 15,655 diagnoses by 101 radiologists based on 155 mammograms (28); (ii) a skin cancer dataset, comprising 4,320 diagnoses by 40 dermatologists based on 108 dermoscopic images of skin lesions (29); (iii) a fingerprint recognition dataset, comprising 1,584 evaluations by 36 professional fingerprint examiners (whom answered whether a pair of fingerprints was matching or not) on 48 fingerprint pairs (30), (iv) a geopolitical forecasting dataset from the Good Judgment Project, containing 8,258 forecasts by 89 forecasters of 94 geopolitical events (31); and (v) a dataset on general knowledge questions (here: which of two cities is larger), containing 99,000 decisions by 99 individuals on 1,000 questions (32). For the medical datasets, the patient's actual health state (i.e. cancer present vs. absent) was known from follow-up research. Similarly, for the forecasting dataset, the correctness of the forecasts was determined from follow-up research (see Materials and Methods for descriptions of all datasets). We use all datasets to investigate the consequences of pooling decision for the variation in accuracy. Response bias, however, can only be studied in three of the datasets (breast cancer, skin cancer and finger recognition) as in the forecasting and the general knowledge dataset only one type of error is possible. Figures 3A-C and 4A, B and *SI Appendix, Fig. S3A-C*, show that in all five datasets individuals differ substantially in their performance characteristics, *SI Appendix, Fig. S4* shows that in the breast cancer, skin cancer, and fingerprint recognition dataset, individuals also differ substantially in their response bias.

To investigate how variation in accuracy and response bias change with group size, within each dataset, we randomly and repeatedly sample (without replacement) two groups of size n (1, 3, 5, 7 and 9). For each draw of two groups, we calculate the performance that both groups achieve across all cases under the majority rule and – as a measure of variation in performance between groups – the absolute difference in performance between both groups. We calculate performance as sensitivity, specificity, and d' for breast and skin cancer, and fingerprint recognition, and as overall accuracy for geopolitical forecasting and general knowledge (see *Materials and Methods* for calculation of d'). For breast and skin cancer, and fingerprint recognition, we also calculate the criterion value and – as a measure of variation in response bias between groups – the absolute difference in criterion between both groups. Within each dataset, for each group size, we repeat this procedure 1,000 times (i.e. we independently sample two groups of a given size) and

determine the mean values of all above measures across these 1,000 simulations. We highlight that we here deliberately use a measure of variation that compares two groups (i.e., absolute difference) in order to ensure that the compared groups always consist of different raters.

Figures 3D-U and 4C-F show the results of this analysis. As predicted, the mean absolute difference (MAD) in sensitivity (Fig. 3D, F, H) and specificity (Fig. 3J, L, N) decreases with increasing group size in breast and skin cancer, and fingerprint recognition. This reduction is substantial in all considered scenarios and strong reductions already occur at relatively small group sizes. For example, the MAD in sensitivity between two randomly selected single experts is 0.125 for breast cancer, 0.13 for skin cancer, and 0.17 for fingerprint recognition; pooling the decisions of five experts reduces these values to 0.06, 0.06, and 0.10—relative reductions of 52%, 54% and 41%, respectively. Importantly, and as reported before (14, 30, 33) mean sensitivity (Fig. 3E, G, I) and specificity (Fig. 3K, M, O) increase with increasing group size in all three contexts. *SI Appendix, Fig. S3D-I* shows that the same patterns are found for the performance measure *d'*. Similarly, in the geopolitical forecasting and general knowledge dataset, increasing group size substantially reduces the MAD in accuracy between groups (Fig. 4C, D) while increasing mean accuracy (Fig. 4E, F). *SI Appendix, Fig. S5* shows that we obtain similar results when using the continuous probability scale in the forecasting dataset (rather than the binary yes/no scale).

Next to variation in accuracy, as predicted, also the MAD in criterion value (i.e. response bias) decreases with increasing group size in breast and skin cancer, and fingerprint recognition (Fig. 3P, R, T). Again, this reduction in MAD in criterion value is substantial in all three datasets and strong reductions already occur at relatively small group sizes. The mean criterion value either remains the same, or slightly in- or decreases with group size (Fig. 3Q, S, U). Thus, as predicted, pooling independent decisions robustly reduces differences in accuracy and response bias between decision-making agents.

So far, we have focused on differences between decision-making agents across a large number of cases. We now ask if and to what extent collective decision-making systems based on the pooling of independent decisions can also reduce variation on the level of the individual case, be it mammogram, skin lesion, fingerprint set, forecasting question, or general knowledge question. To this end, we randomly and repeatedly sample two individuals and determine whether or not both individuals give the same response to each individual case within each dataset. This is repeated 2,500 times per case, and we then calculate

the average frequency of agreement per case. We use the same procedure for groups: for each case within each dataset, we randomly and repeatedly sample two groups of *n* individuals (5 and 9), and determine whether or not these two groups arrive at the same response under the majority rule. This is repeated 2,500 times per case per group size, and we then calculate the average frequency of agreement per case per group size.

Figure 5 shows the results of this analysis. In all datasets, there is a relatively high frequency of cases for which there is a high chance that two randomly sampled individuals disagree. To illustrate, in the case of breast cancer, 71 out of 155 cases have an agreement level below 0.6, implying that for each of these cases, there is an at least 40% chance that two randomly sampled radiologists disagree. Importantly, with increasing group size, in all datasets, this distribution shifts to the right: pooling decisions thus systematically decreases the number of cases where decision-making agents disagree and increases the number of cases where they agree, illustrating that pooling decisions also reduces outcome variation between decision-making agents at the case level.

Discussion

Collective decision-making systems based on pooling independent decisions can be a powerful approach to increase decision accuracy (23, 34-37). But it can do even more. Using a combined theoretical and data-driven approach, we analyzed an important but up to now largely neglected performance dimension of collective vs. individual decision-making systems, the variation in outcomes between different decision-making agents. Our theoretical and empirical results arrive at the same conclusion: The pooling of independent judgments is a powerful pathway to reduce outcome variation between decision-making agents. As has routinely been demonstrated, individual experts differ substantially in both their accuracy level and response bias (e.g. 9, 14-18, 19 and Fig. 3A-C and SI Appendix, Fig. S4). Our results suggest that this undesirable variation can be substantially reduced by a system that combines independent decisions of experts.

A key implication of our findings is that collective decision-making systems based on pooling independent decisions will often be more reliable and predictable than systems based on individual decision

makers. While this is important in itself, in many contexts this can be expected to have major consequences for the actual and perceived fairness of a given system (38, 39). In judicial decision-making, for example, judges who differ in their response bias (i.e. how much evidence they require to convict a suspect), are considered detrimental to the principle of equal treatment under the law. Beyond boosting fairness, more reliable and predictable decision-making systems, with less outcome variation between agents, are typically also perceived as more trustworthy, collective decision-making systems based on pooling decisions may thus be an important step towards preventing the erosion of institutional trust.

Response biases play a crucial role in decision making in a wide range of domains, including medical diagnostics and judicial, political and economic decision-making (8, 21, 40). A key normative question here is how much evidence should be enough to classify a mammogram as malignant, or to convict a defendant (19, 41)? Individual decision makers differ with respect to that decision threshold (i.e., response bias), which contributes to the differences in decisions outcomes among decision makers. Collective systems based on pooling independent decisions can be a powerful corrective to such unwanted variation. Furthermore, even in contexts that lack a clear normatively defensible response bias, it may be desirable to favor decision accuracy in one state over the other. For example, medical patients often differ in their preferences (42, 43) and taking such individual-specific preferences into account can thus be important when providing treatment recommendations. Also for such individual-tailored approaches, the system of pooling independent judgments can be useful. Specifically, while we have focused in our analyses on the majority rule, independent decisions can also be aggregated with the more flexible quorum thresholds that allow the fine-tuning of decisions under any error cost scheme (8, 24). In (33), we illustrate this approach in the context of skin cancer diagnostics.

The importance of variation in forecasts, judgments and decisions has been acknowledged in several areas across the behavioural sciences (e.g. 44, 45, 46). For example, in the literature on optimal portfolio selection, it is well-known that it is important to take into account both the mean rate and the variance (i.e., risk) of returns for securities (47); also in the literature on forecasting (48, 49) and machine learning (50), it has been highlighted that aggregation reduces risk. Similarly, in animal behaviour, the theory on risk-sensitive foraging is centered on the insight that – next to mean foraging returns – variation

in foraging returns over time is a key fitness determinant (51). It will be interesting to investigate whether our findings may also have bearings in these contexts.

One of the future questions is whether other systems of collective decision-making like, for example, interacting individuals that discuss with each other, also achieve higher levels of between-group agreement. It is well known that directly interacting individuals can, under some circumstances, achieve higher accuracy than individual decision makers and/or systems based on pooling independent judgments (52). Little is known, however, about the consequences that collective systems based on interacting groups have for variation in outcomes between different decision-making agents (i.e. interacting groups). While pooling independent judgments is firmly rooted in statistical principles, the dynamics of interacting groups are governed by additional principles relevant for interaction and communication (e.g. what pieces of information are mentioned during a discussion) (53, 54). For example, cognitive strategies, such as a confirmation strategy, may amplify when individuals with similar biases (e.g. in terms of criterion value) interact. In mock juries it has been observed that deliberation in groups (but not as an individual) increases the leniency of sentences (55). Whether and how collective systems based on interacting groups affect variation in outcomes between groups is thus not straightforward and needs substantial theoretical and empirical scrutiny.

Over the last decades, researchers across the behavioural sciences have put massive effort into mapping the relative advantages and disadvantages of individual vs. collective decision-making systems in terms of accuracy. Next to accuracy, variation in outcomes between decision-making agents is a key benchmark of decision-making systems. As our results demonstrate, collective decision-making systems based on pooling independent decision are a powerful approach to reduce such variation, thus promoting reliability, predictability, fairness, and, possibly, even trust.

Materials and Methods

Breast cancer dataset

The full information on the breast cancer dataset can be found in (28), and we summarized the dataset in (9, 13, 14). Therefore, we here provide a brief summary largely adopting these earlier descriptions. Mammograms were randomly selected from screening examinations performed on women aged 40–69 between 2000 and 2003 from US mammography registries affiliated with the Breast Cancer Surveillance Consortium. Radiologists who interpreted mammograms at facilities affiliated with these registries between January 2005 and December 2006 were invited to participate in this study, as were radiologists from Oregon, Washington, North Carolina, San Francisco, and New Mexico. Of the 409 radiologists invited, 101 completed all procedures and were included in the data analyses. Each screening examination included images from the current examination and one previous examination (allowing the radiologists to compare potential changes over time) and presented the craniocaudal and mediolateral oblique views of each breast (four views per woman for each of the screening and comparison examinations). This approach is standard practice in the United States. Women who were diagnosed with cancer within 12 months of the mammograms were classified as cancer patients (n = 27). Women who remained cancer-free for a period of two years were classified as noncancerous patients (n = 128; 17% prevalence).

Radiologists viewed the digitized images on a computer (home computer, office computer, or laptop provided as part of the original study). All computers were required to meet all viewing requirements of clinical practice, including a large screen and high-resolution graphics (≥1,280 × 1,024 pixels and a 1280MB video-card with 32-bit color). Radiologists saw two images at the same time (left and right breasts) and were able to alternate quickly (≤1 s) between paired images, to magnify a selected part of an image, and to identify abnormalities by clicking on the screen. Each case presented craniocaudal and mediolateral oblique views of both breasts simultaneously, followed by each view in combination with its prior comparison image. Cases were shown in random order. Radiologists were instructed to diagnose them using the same approach they used in clinical practice (i.e. using the breast imaging reporting and data system lexicon to classify their diagnoses, including their decision that a woman be recalled for further examination). Radiologists evaluated the cases in two stages. In stage 1, four test sets were created, each

containing 109 cases. Radiologists were randomly assigned to one of the four test sets. In stage 2, one test set containing 110 cases was created and presented to all radiologists. Some of the cases used in stage 2 had already been evaluated by some of the radiologists in stage 1. To avoid having the same radiologist evaluate a case twice, we excluded all cases from stage 2 that had already been viewed by that radiologist in stage 1. Moreover, we only included cases present in all four test sets in order to ensure that each radiologist evaluated the same set of cases, resulting in 155 unique cases. Between the two stages, radiologists were randomly assigned to one of three intervention treatments. Because there were no strong treatment differences (56), we pooled the data from stages 1 and 2. In our analysis, we treated the recommendation that a woman should be recalled for further examination as a positive test result.

Skin cancer dataset

The full information on the skin cancer dataset can be found in (29), and we summarized the dataset in (9, 13, 33). Therefore, we here provide a brief summary largely adopting these earlier descriptions. This dataset comprises 4,320 diagnoses by 40 dermatologists of 108 skin lesions and was collected during a web-based consensus meeting. Skin lesions were obtained from the Department of Dermatology, University Frederico II (Naples, Italy); the Department of Dermatology, University of L'Aquila (Italy); the Department of Dermatology, University of Graz (Austria); the Sydney Melanoma Unit, Royal Prince Alfred Hospital (Camperdown, Australia); and Skin and Cancer Associates (Plantation, Florida). The study was designed to diagnose whether or not a skin lesion was a melanoma, the most dangerous type of skin cancer. Histopathological specimens of all skin lesions were available and judged by a histopathology panel (melanoma: $n = 27$, no melanoma: $n = 81$; 25% prevalence). All dermatologists that participated in the study had a minimum of five years of experience in dermoscopy practice, teaching, and research. Dermatologists first received a training procedure in which they familiarized themselves with the study's definitions and procedures in web-based tutorials with 20 sample skin lesions. They subsequently evaluated 108 skin lesions in a two-step online procedure. First, they used an algorithm to differentiate melanocytic from nonmelanocytic lesions. Whenever a lesion was evaluated as melanocytic, dermatologists were asked to classify it as either melanoma or a benign melanocytic lesion, using four different algorithms. We use the diagnostic algorithm with the highest diagnostic accuracy, which is also the one most widely used for

melanoma detection: pattern analysis. We treated the decision to classify a lesion as melanoma as a positive test result.

Fingerprint analyses

The full information on the fingerprint analyses dataset can be found in (30), therefore, we provide a brief summary only. The dataset comprises 1,728 evaluations by 36 professional fingerprint examiners on 48 fingerprint pairs. The fingerprint examiners were recruited from the Australian Federal Police, Queensland Police Service, Victoria Police, and New South Wales Police (mean experience = 16.4 years). Also novices participated in testing, but we only used data from the professional examiners. Each of the 36 fingerprint examiners was presented with the same set of 24 fingerprint pairs from the same finger (targets) and 24 highly similar pairs from different fingers (distractors) in a different random order. Each pair consisted of a crime-scene "latent" fingerprint and a fully-rolled "arrest" fingerprint, and participants were asked to provide a rating on a 12-point scale ranging from 1 (Sure Different) to 12 (Sure Same). The distractors were created by running each latent fingerprint through the National Australian Fingerprint Identification System—which consists of roughly 67 million fingerprints—to return the most similar exemplars from the database. On the first 44 of 48 trials (22 targets, 22 distractors), participants were given 20 seconds to examine the prints. On the final four trials (2 targets, 2 distractors), they had an unlimited amount of time to make a decision. For consistency, we excluded these last four trials, resulting in a total of 1,584 evaluations by 36 professional fingerprint examiners on 48 fingerprint pairs. To investigate the case of binary decision-making, we converted the answers from the 12-point scale into 'different' (6 and below) and 'same' (7 and above).

Forecasting dataset

The forecasting dataset is part of the Good Judgment Project (31) and we summarized the dataset in (13). Therefore, we provide a brief summary largely adopting this earlier description. The Good Judgement Project is a large-scale forecasting project, running over several years and using a wide variety of participants and settings (e.g., training schedules, team competitions). Participants were free to enter the forecasting competition, and the subject pool consisted of a mix of laypeople and geopolitical experts. We used data from the first year of the forecasting project (57). In this year, 102 questions, such as "Will Serbia

be officially granted EU candidacy by 31 December 2011?" and "Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?" had to be forecasted. Participants were asked to estimate the probability of the future event, on a scale from 0 to 1. We only included data from the individual condition (i.e. we excluded individuals who observed crowd information or participated in prediction markets). We excluded questions with more than two possible answers and questions for which the correct answer could not be irrefutably determined (n = 8), resulting in 94 questions. Finally, we excluded forecasters who answered less than 90 questions, resulting in 89 forecasters. The total dataset we used contained 8,258 forecasts by 89 forecasters on 94 geopolitical events. Sometimes, forecasters updated their forecasts over time, thereby giving multiple responses. In such cases, we used their first forecast only. To investigate the case of binary decision-making, we converted the probability scores into 0 (probabilities < 0.5) and 1 (probabilities > 0.5). 0.5 scores were randomly converted to either 0 or 1. In *SI Appendix, Fig. S5*, we investigate a scenario when using the probability forecasts directly without any conversion, using the Brier score as accuracy measure.

General knowledge dataset

This dataset is based on Study 3 in (32) and we summarized the dataset in (13). Therefore, we here provide a brief summary largely adopting the earlier description. The dataset contains binary responses to the question, "Which of the following two cities has more inhabitants?" The stimulus set consisted of 1,000 randomly generated pairs of cities from a list of the 100 most populous cities in the United States (US) in 2010 as determined by the US Census Bureau. After seeing a fixation cross, participants observed a pair of cities and after 1.6 seconds, they were cued to decide. Participants rated 1,000 pairs in two sessions. Participants (n = 109) were recruited from the Michigan State University (MSU) psychology research participant pool and received class credits plus a $0–$4 bonus per session. We excluded participants who did not complete both sessions, resulting in 99 participants, all of whom provided a decision on each of the 1,000 city pairs.

Ethics statement and data availability

The breast cancer data were assembled at the BCSC Statistical Coordinating Center (SCC) in Seattle and analyzed at the Max Planck Institute for Human Development (MPIB), Germany. Each registry, the SCC and the MPIB, received institutional review board approval for active and passive consent processes or were granted a waiver of consent to enroll participants, pool data, and perform statistical analysis. All procedures were in accordance with the Health Insurance Portability and Accountability Act. All data were anonymized to protect the identities of women, radiologists, and facilities. The BCSC holds legal ownership of the data. Information regarding data requests can be found at bcsc-research.org/. For the skin cancer data, the review board of the Second University of Naples waived approval because the study did not affect routine procedures. All participating dermatologists signed a consent form before participating in the study. The skin cancer dataset can be accessed at pnas.org/content/113/31/8777/tab-figures-data. The fingerprint recognition study was cleared by the ethical board of The University of Queensland and The University of Adelaide. The fingerprint dataset can be accessed at osf.io/hgx3s. The geopolitical dataset is part of the Good Judgment Project and accessible at dataverse.harvard.edu/dataverse/gjp. Participants in the general knowledge task gave informed consent according to the MSU Institutional Review Board guidelines. The general knowledge dataset is accessible at osf.io/cuzqm/. The code for reproducing the results and figures of the statistical argument (Fig. 1) and the numerical simulations (Fig. 2) are uploaded at the Open Science Framework: link. Also the code for reproducing the results and figures of the skin cancer dataset (Fig. 3 and Fig. 5) are uploaded to illustrate how the results were calculated for one of the datasets. This link is for Reviewer's inspection and will be made public after manuscript acceptance.

Calculation of discrimination ability and criterion

For the simulations of the empirical data, we repeatedly and randomly drew groups of different sizes. The decisions of the drawn individuals were combined using the majority rule. Based on individual or group decisions, we calculated the number of hits (H), misses (M), false alarms (FA) and correct rejections (CR). From this we calculated the hit rate (HR) as HR = H / (H + M) and the false-alarm rate (FAR) as FAR = FA / (FA + CR). From this we calculated discrimination ability d' and criterion c using:

$d' = z(HR) - z(FAR)$ and

$c = -(z(HR) + z(FAR))/2$.

where z transformation converts the hit and false-alarm rate to a z-score (i.e. to standard deviation units). Note that we followed the standard approach to add 0.5 to each of the four categories (i.e. hits, misses, FA, CR) in order to avoid the possibility of infinite values (22).

Acknowledgements

We thank Jose Cayere, Amy Buzby, and the American College of Radiology for their technical assistance in developing and supporting the implementation of the test sets; the expert radiologists Larry Bassett, Barbara Monsees, Ed Sickles, and Matthew Wallis; and the participating women, facilities, and radiologists for the data they provided. The BCSC investigators are listed at www.breastscreening.cancer.gov/. This work was supported by the American Cancer Society using a donation from the Longaberger Company's Horizon of Hope Campaign (Grants SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRSG-06-281, SIRSG-09-270-01, SIRSG-09-271-01, and SIRSG-06-290-04); by the Breast Cancer Stamp Fund; and by the National Cancer Institute Breast Cancer Surveillance Consortium (Grant HHSN261201100031C). The collection of cancer and vital status data used in this study was supported, in part, by several state public health departments and cancer registries throughout the United States. A full description of these sources is provided at www.breastscreening.cancer.gov/work/acknowledgement.html. This work was further funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

References

1. Conradt L & List C (2009) Group decisions in humans and animals: a survey. *Philos T Roy Soc B* 364(1518):719-742.
2. Bang D & Frith CD (2017) Making better decisions in groups. *R Soc Open Sci* 4(8):170193.
3. Kerr NL & Tindale RS (2004) Group performance and decision making. *Annu Rev Psychol* 55:623-655.
4. Gully SM, Incalcaterra KA, Joshi A, & Beaubien JM (2002) A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *J Appl Psychol* 87(5):819-832.
5. El Zein M, Bahrami B, & Hertwig R (2019) Shared responsibility in collective decisions. *Nature human behaviour* 3(6):554-559.
6. Bahrami B*, et al.* (2010) Optimally interacting minds. *Science* 329(5995):1081-1085.
7. Clément RJG*, et al.* (2013) Collective cognition in humans: Groups outperform their best members in a sentence reconstruction task. *PLoS ONE* 8(10):e77943.
8. Wolf M, Kurvers RHJM, Ward AJW, Krause S, & Krause J (2013) Accurate decisions in an uncertain world: Collective cognition increases true positives while decreasing false positives. *Proc R Soc Lond B* 280(1756):20122777.
9. Kurvers RHJM*, et al.* (2016) Boosting medical diagnostics by pooling independent judgments. *Proc. Natl. Acad. Sci. U. S. A.* 113(31):8777-8782.
10. Lorenz J, Rauhut H, Schweitzer F, & Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *P Natl Acad Sci USA* 108(22):9020-9025.
11. Woolley AW, Chabris CF, Pentland A, Hashmi N, & Malone TW (2010) Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330(6004):686-688.
12. Koriat A (2012) When are two heads better than one and why? *Science* 336(6079):360-362.
13. Kurvers RH*, et al.* (2019) How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances* 5(11):eaaw9011.
14. Wolf M, Krause J, Carney PA, Bogart A, & Kurvers RHJM (2015) Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLoS ONE* 10(8):e0134269.
15. Koran LM (1975) The reliability of clinical methods, data and judgments. *N. Engl. J. Med.* 293(13):642-646.
16. Burgman MA (2016) *Trusting judgements: how to get the best out of experts* (Cambridge University Press).
17. O'Sullivan M & Ekman P (2004) 12 The wizards of deception detection. *The detection of deception in forensic contexts*:269.
18. Mellers B*, et al.* (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3):267-281.

19. Deneef P & Kent DL (1993) Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. *Med. Decis. Making* 13(2):126-132.
20. Hammond KR (1996) *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice* (Oxford University Press on Demand).
21. DeKay ML (1996) The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law & Social Inquiry* 21(1):95-132.
22. Macmillan NA & Creelman CD (2005) *Detection theory: A user's guide* (2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ).
23. Grofman B, Owen G, & Feld SL (1983) Thirteen theorems in search of the truth. *Theor Decis* 15(3):261-278.
24. Marshall JA, Kurvers RH, Krause J, & Wolf M (2019) Quorums enable optimal pooling of independent judgements in biological systems. *Elife* 8:e40368.
25. Sorkin RD, Hays CJ, & West R (2001) Signal-detection analysis of group decision making. *Psychol Rev* 108(1):183-203.
26. Sorkin RD & Dai H (1994) Signal detection analysis of the ideal group. *Organ. Behav. Hum. Decis. Process.* 60(1):1-13.
27. Macmillan NA & Creelman CD (1990) Response bias: Characteristics of detection theory, threshold theory, and" nonparametric" indexes. *Psychological Bulletin* 107(3):401.
28. Carney PA*, et al.* (2012) Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *Am. J. Roentgenol.* 198(4):970-978.
29. Argenziano G*, et al.* (2003) Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *J. Am. Acad. Dermatol.* 48(5):679-693.
30. Tangen JM, Kent K, & Searston RA (2020) Collective Intelligence in Fingerprint Analysis. *Cognitive Research: Principles and Implication* 5(23).
31. Ungar L, Mellers B, Satopää V, Tetlock P, & Baron J (2012) The good judgment project: A large scale test of different methods of combining expert predictions. *AAAI Fall Symposium Series*.
32. Yu S, Pleskac TJ, & Zeigenfuse MD (2015) Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* 144(2):489.
33. Kurvers RHJM, Krause J, Argenziano G, Zalaudek I, & Wolf M (2015) Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol* 151(12):1-8.
34. Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Knopf Doubleday Publishing Group).
35. Herzog SM, Litvinova A, Yahosseini KS, Novaes Tump A, & Kurvers RHJM (2019) The ecological rationality of the wisdom of crowds. *Taming Uncertainty*, (MIT Press, Cambridge, Massachusetts).
36. Mannes AE, Soll JB, & Larrick RP (2014) The wisdom of select crowds. *J Pers Soc Psychol* 107(2):276.

37. Hastie R & Kameda T (2005) The robust beauty of majority rules in group decisions. *Psychol Rev* 112(2):494-508.
38. Tyler T, Degoey P, & Smith H (1996) Understanding why the justice of group procedures matters: A test of the psychological dynamics of the group-value model. *J Pers Soc Psychol* 70(5):913.
39. Wood G, Tyler TR, & Papachristos AV (2020) Procedural justice training reduces police use of force and complaints against officers. *P Natl Acad Sci USA* 117(18):9815-9821.
40. Swets JA, Dawes RM, & Monahan J (2000) Psychological Science Can Improve Diagnostic Decisions. *Psychol Sci Public Interest* 1(1):1-26.
41. Swets JA (1992) The science of choosing the right decision threshold in high-stakes diagnostics. *Am. Psychol.* 47(4):522-532.
42. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, & Welch HG (2000) US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *BMJ* 320(7250):1635-1640.
43. O'Connor AM, Légaré F, & Stacey D (2003) Risk communication in practice: the contribution of decision aids. *BMJ* 327(7417):736-740.
44. Stewart TR (2001) Improving reliability of judgmental forecasts. *Principles of forecasting*, (Springer), pp 81-106.
45. Kahneman D, Rosenfield A, Gandhi L, & Blaser T (2016) Noise. *Inconsistent Decision-making. HBR*:38-46.
46. Litvinova A, Kurvers RH, Hertwig R, & Herzog SM (2019) When experts make inconsistent decisions.
47. Brealey RA, Myers SC, Allen F, & Mohanty P (2012) *Principles of corporate finance* (Tata McGraw-Hill Education).
48. Hibon M & Evgeniou T (2005) To combine or not to combine: selecting among forecasts and their combinations. *International journal of forecasting* 21(1):15-24.
49. Lichtendahl Jr KC & Winkler RL (2020) Why do some combinations perform better than others? *International Journal of Forecasting* 36(1):142-149.
50. Kuncheva LI (2014) *Combining pattern classifiers: methods and algorithms* (John Wiley & Sons).
51. McNamara JM & Houston AI (1992) Risk-sensitive foraging: a review of the theory. *Bull Math Biol* 54(2-3):355-378.
52. Navajas J, Niella T, Garbulsky G, Bahrami B, & Sigman M (2018) Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2(2):126-132.
53. Stasser G & Abele S (2020) Collective Choice, Collaboration, and Communication. *Annu Rev Psychol* 71(1):589-612.
54. Stasser G & Titus W (1985) Pooling of unshared information in group decision making: Biased information sampling during discussion. *J Pers Soc Psychol* 48(6):1467.

55. MacCoun RJ & Kerr NL (1988) Asymmetric influence in mock jury deliberation: Jurors' bias for leniency. *J Pers Soc Psychol* 54(1):21.
56. Geller BM*, et al.* (2014) Educational interventions to improve screening mammography interpretation: A randomized controlled trial. *Am. J. Roentgenol.* 202(6):W586-W596.
57. Anonymous (2016) Good Judgment Project. GJP Data. Harvard Dataverse, V1.

Figures

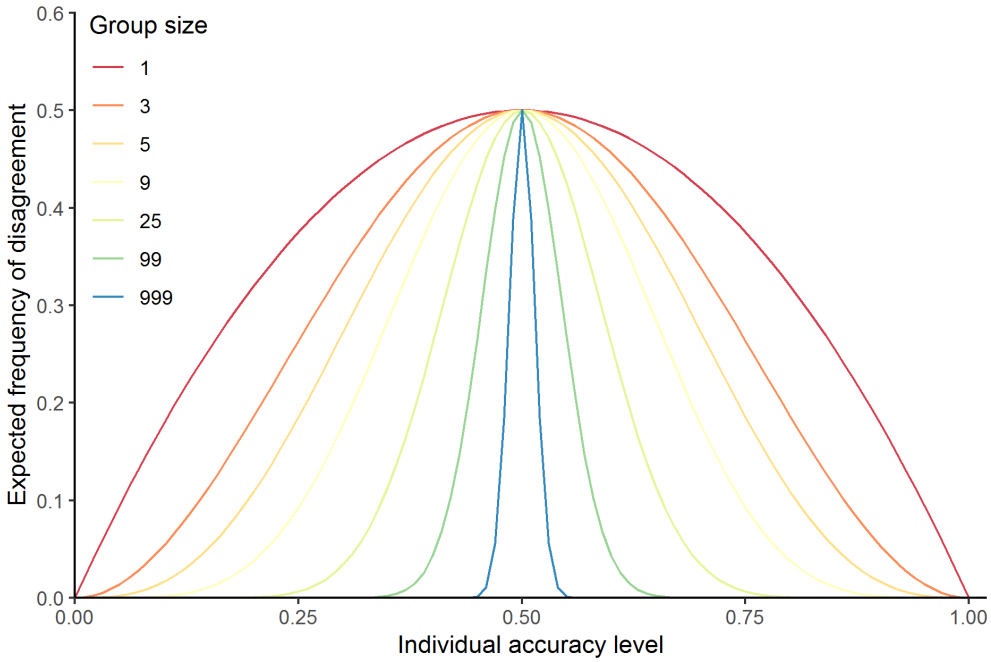

Figure 1. Statistical argument: Pooling independent decisions reduces (stochastic) outcome variation. We here consider the most basic scenario where individual decision makers do not differ in accuracy or response bias. While all individuals achieve the same accuracy level, they may still differ (stochastically) in how they respond to specific cases. The expected frequency of disagreement between two decision-making agents (i.e. two individuals, two groups pooling independent decisions with a majority rule) can be calculated from the binomial distribution (see Results). For any level of individual accuracy (x-axis), the expected frequency of cases where two agents disagree is shown. As can be seen – with the exception of accuracies 0.0, 0.5 and 1.0 – pooling decisions is predicted to systematically reduce the frequency of cases with disagreement between agents, with larger groups achieving larger reductions than smaller groups. This reduction can be explained by (i) the link between accuracy levels and the expected frequency of disagreement and (ii) the effect of pooling decisions on accuracy levels: as long as individual decision maker achieve an accuracy level above chance (i.e. >0.5), pooling decisions increases accuracy levels which, in turn, decreases the expected disagreement (an analogous argument holds for individual accuracy levels <0.5, which are shown for completeness, see Results for details). Only odd group sizes are calculated to avoid the need for a tie-breaking rule.

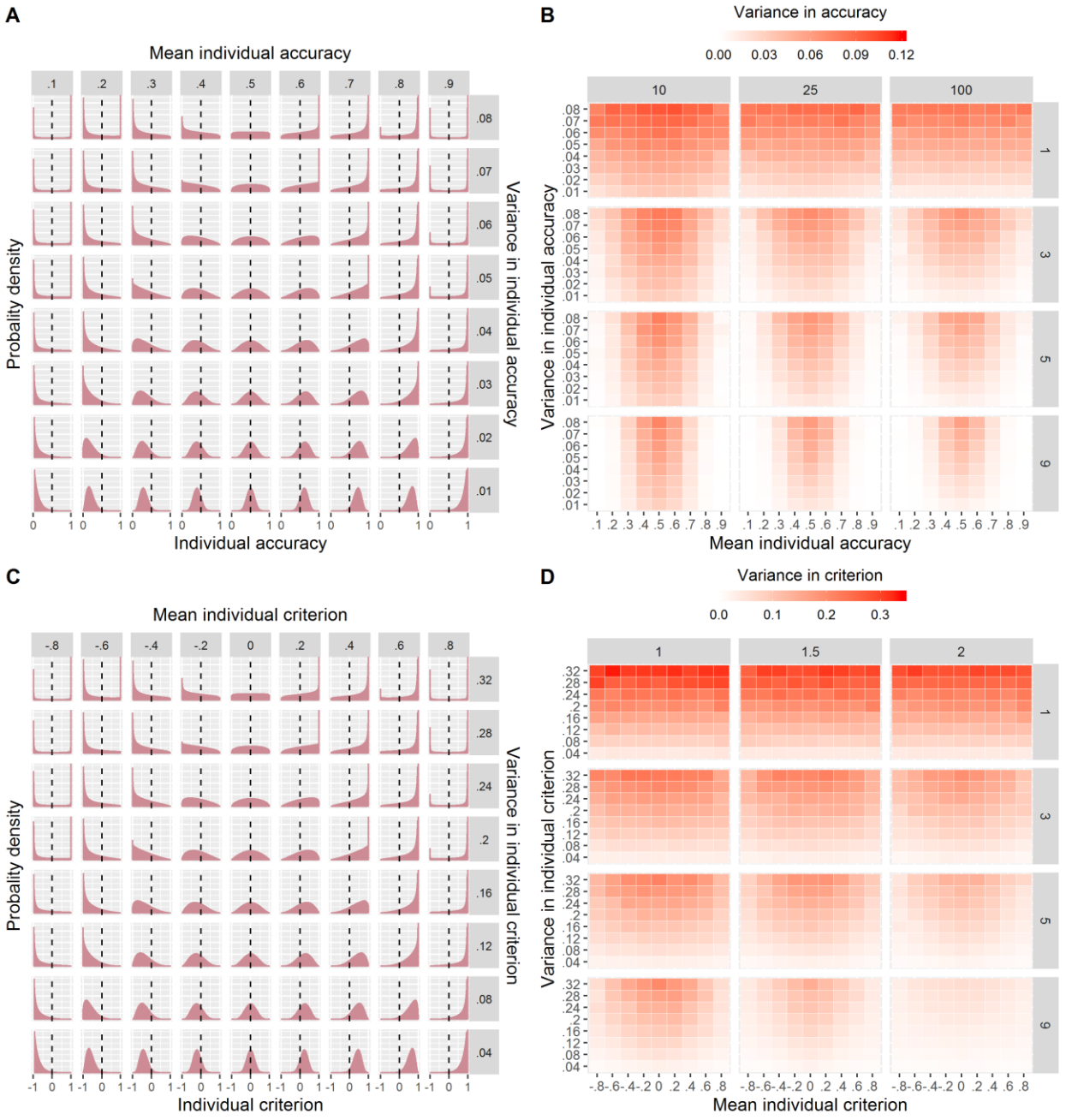

Figure 2. Numerical simulations: Pooling independent decisions reduces variation in accuracy and response bias. (A) For the numerical simulations, we sampled decision makers from a wide range of populations of decision makers differing in their performance distribution (x-axis: individual accuracy; y-axis: probability density). We created those by systematically varying the mean (values on top) and variance (values on the right) of the beta distribution. Dashed vertical lines indicate chance level of raters (i.e. accuracy of 0.5). (B) The variance in accuracy between individuals/groups for differently-sized groups (*n* =

1, 3, 5, and 9; subpanel rows), making m = 10, 25, and 100 decisions (subpanel columns). Within each subpanel, the tiles correspond to raters drawn from the population (i.e. accuracy distribution) of the associated tiles (i.e. mean-variance combination) in (A). Within each subpanel, the middle column (i.e. mean accuracy = 0.5) corresponds to rater populations performing on average at chance level, with tiles to the left (right) of that column indicating populations performing on average below (above) chance. Increasingly red colors indicate increasing variance in accuracy between different individuals, or groups employing a majority rule. For each unique combination of group size n, decision cases m, and accuracy distribution, the shown variance corresponds to the variance between 10,000 independently sampled groups. As can be seen, independent of the specific accuracy distribution and number of decision cases considered, we robustly observe that increasing group size reduces the variance in accuracy between groups. This effect is smallest for populations at chance level, and increases the closer the mean accuracy moves towards 1 or 0. (C) To investigate response bias, we sampled decision makers from a wide range of populations of decision makers differing in their criterion value (x-axis: criterion parameter; y-axis: probability density). We created those by systematically varying the mean (values on top) and variance (values on the right) of the beta distribution, where we transformed the beta range from [0,1] to [-1,1] to achieve a broader range of criterion values. Dashed vertical lines indicate no response bias (i.e. criterion value = 0), raters increasingly to the left (or right) of that line correspond to raters that put increasingly more weight on accuracy in state 1 or state 2, respectively. Note that within any given population, we assumed that all individuals are characterized by the same discrimination ability. (D) The variance in response bias between individuals/groups for differently-sized groups (n = 1, 3, 5, and 9; subpanel rows) and three different discrimination abilities 1, 1.5 and 2 (subpanel columns). Within each subpanel, the tiles correspond to raters drawn from the population (i.e. criterion distribution) of the associated tiles (i.e. mean-variance combination) in (C). Increasingly red colors indicate increasing variance in response bias between different individuals, or groups employing a majority rule. For each unique combination of group size n, discrimination ability d', and response bias c, we independently sampled 10,000 groups and calculated the variance in response bias between these 10,000 groups. As can be seen, independent of the specific distribution and discrimination ability in the population, we find that increasing group size robustly reduces the variation in response bias between groups.

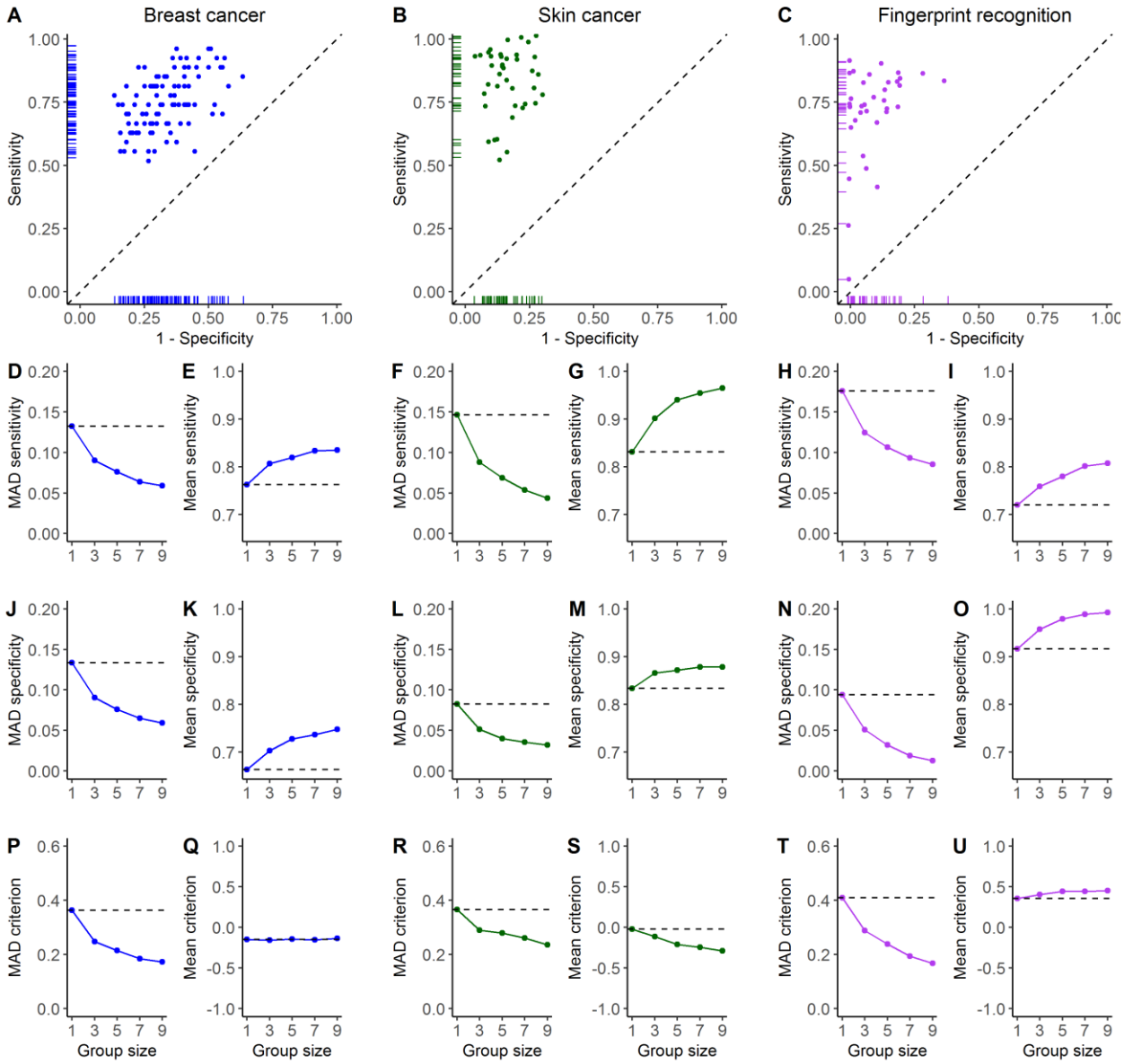

Figure 3. Pooling independent decisions reduces variation in accuracy and response bias in breast and skin cancer detection, and fingerprint recognition. (A-C). True and false positive rates (i.e. sensitivity and 1-specificity, respectively) of each rater in the three datasets. Each dot corresponds to one rater. The dashed diagonal corresponds to those points that can be achieved by a random classifier; dots above (below) the diagonal thus indicate a performance above (below) chance. As can be seen, in all three domains, raters differ substantially in both accuracy components (i.e. sensitivity and specificity). (D-U) For each dataset and each group size, we repeatedly and randomly sampled two groups and – as a measure

of outcome variation between groups – calculated the mean absolute differences (MAD) in sensitivity, specificity and criterion (i.e. response bias) between two randomly sampled groups. As predicted, in all three datasets, compared to the baseline levels of variation between individual decision makers (dashed lines), substantial reductions in variation (i.e. MAD) in (D, F, H) sensitivity, (J, L, N) specificity, and (P, R, T) criterion are achieved when pooling independent decisions; as expected, these reductions increase with increasing group size. Importantly, these reductions in variation are accompanied by an increase in mean (E, G, I) sensitivity and (K, M, O) specificity; mean criterion either (Q) remained unchanged, (S) slightly decreased, or (U) slightly increased.

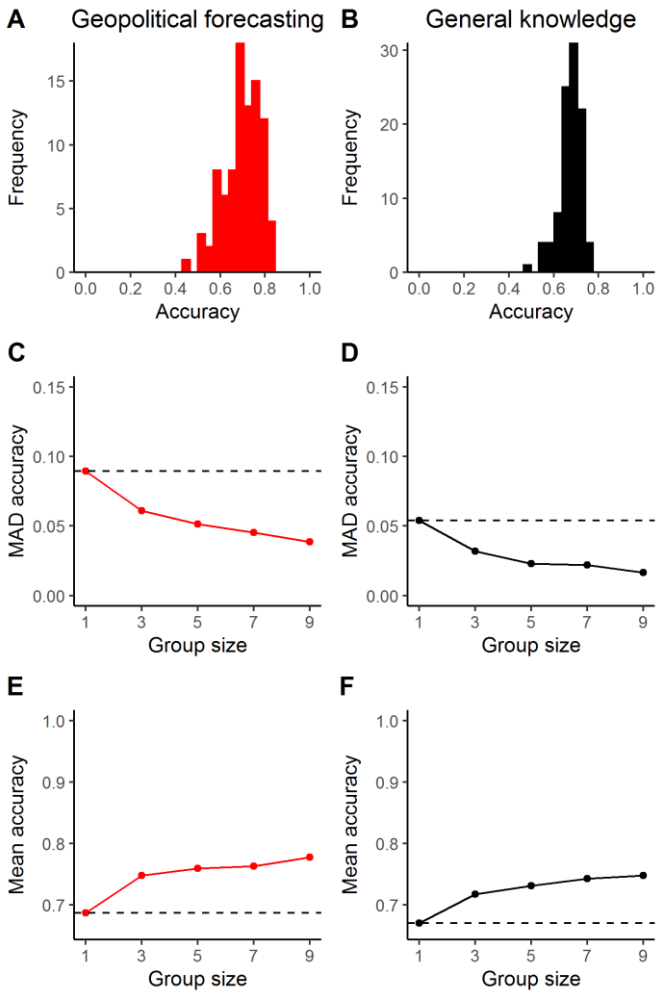

Figure 4. Pooling independent decisions reduces variation in geopolitical forecasting and a general knowledge task. (A, B) Frequency distribution showing the distribution of accuracy levels of raters in the (A) geopolitical forecasting, and (B) general knowledge dataset, showing that raters in both domains differ substantially in accuracy levels. (C, D) For each dataset and each group size (*n* = 1, 3, 5, 7, and 9), we repeatedly and randomly sampled two groups and – as a measure of variation between groups – calculated the mean absolute differences (MAD) in accuracy. As predicted, in both datasets, compared to the baseline level of variation between individual decision makers (dashed lines), substantial reductions in variation (i.e. MAD) in accuracy are achieved when employing collective systems based on pooling independent decisions, as expected, these reductions increase with increasing group size. (E, F) Importantly, these reductions in variation are accompanied by an increased mean accuracy in both datasets.

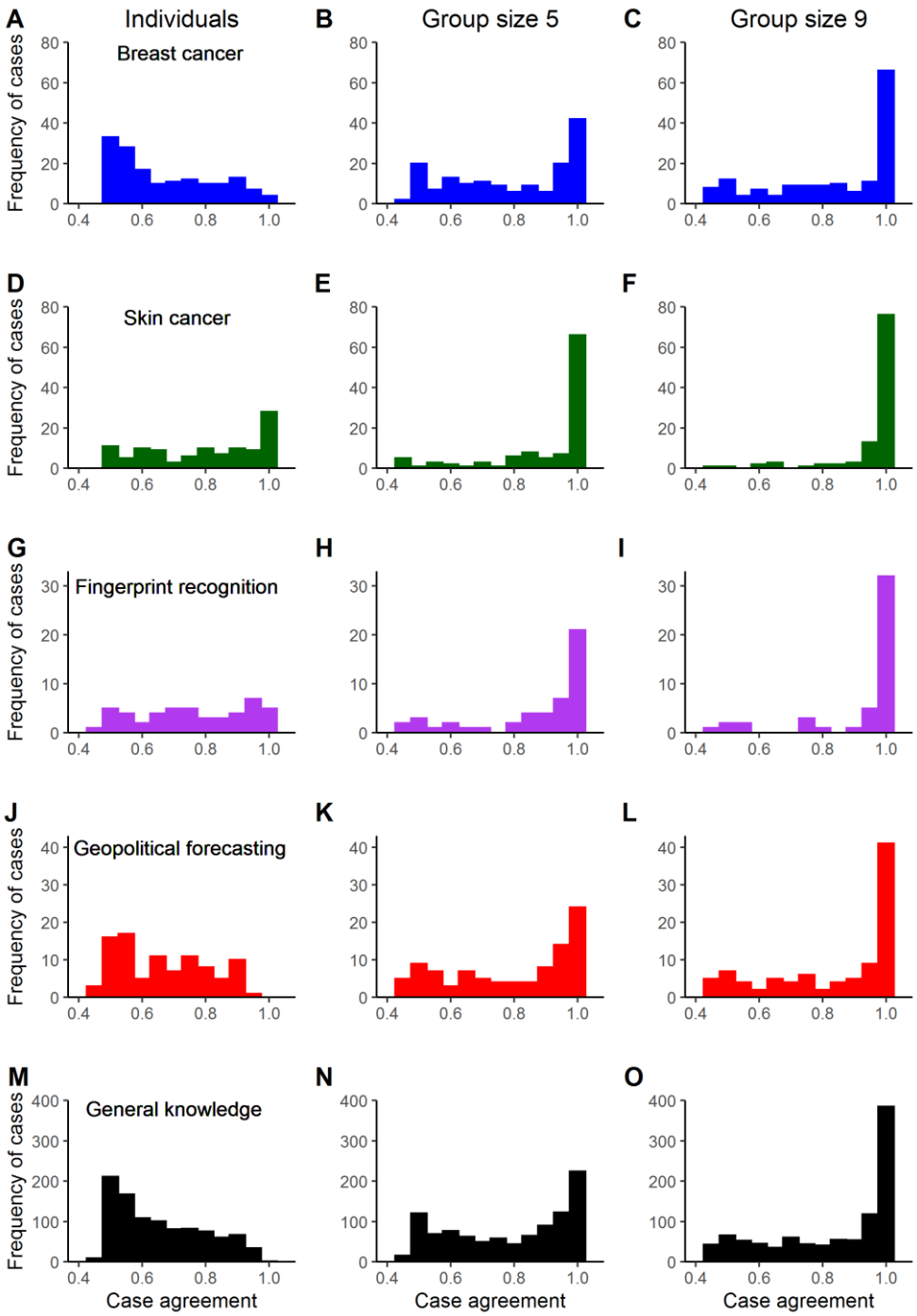

Figure 5. Pooling independent decisions reduces variation at the case level in breast and skin cancer detection, fingerprint recognition, geopolitical forecasting and a general knowledge task. The absolute frequency of cases for which (i) two randomly selected individuals, (ii) two groups each of five randomly selected individuals and (iii) two groups each of nine randomly selected individuals agree, for (A-C) breast cancer, (D-F) skin cancer, (G-I) fingerprint recognition, (J-L) geopolitical forecasting and (M-O) a

general knowledge task. In all five datasets, there is a substantial frequency of cases for which two randomly sampled individuals disagree. Compared to this baseline level, employing a collective system based on pooling independent decisions systematically decreases the number of cases where individuals disagree and increases the number of cases where they agree (i.e. the distributions shift to the right) and, as expected, this effect increased with increasing group size.

Supporting information

Towards fairer and more reliable decision-making systems: pooling decisions decreases variation in accuracy and response bias

Ralf H. J. M. Kurvers, Stefan M. Herzog, Ralph Hertwig, Jens Krause & Max Wolf

SUPPLEMENTARY FIGURES

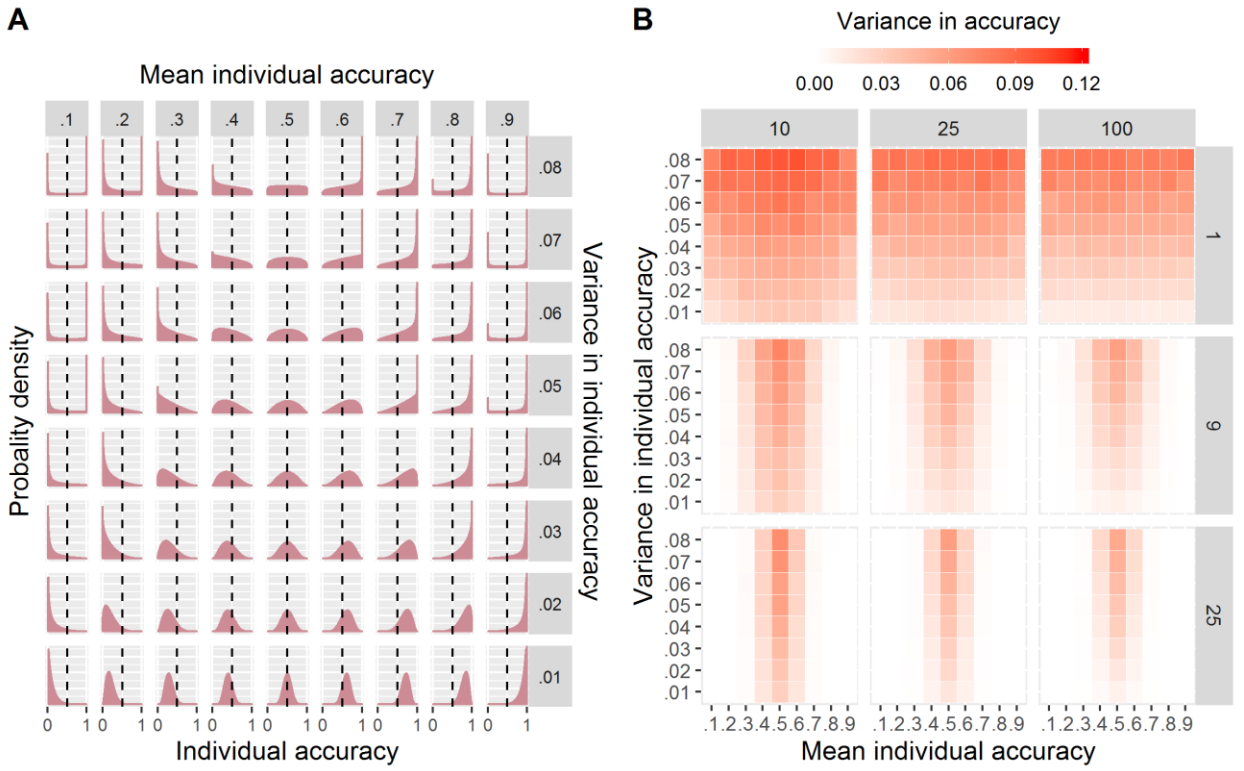

Fig. S1. Numerical simulations: Reduction in variation in accuracy when pooling decision from larger groups. This figure corresponds to Fig. 2A, and 2B in the main text with the exception that in (B) we skipped the results for group size 3 and 5 and added the results for group size 25; individual decision makers (group size 1) and group size 9 were kept as reference levels. As can be seen, compared to decision systems based on pooling from groups of size 9, pooling from groups of size 25 reduces the variance in accuracy between decision-making agents (i.e. groups) even further. In fact, whenever decision makers are sampled from populations with an average accuracy not too close to 0.5, variation between decision makers almost completely vanishes.

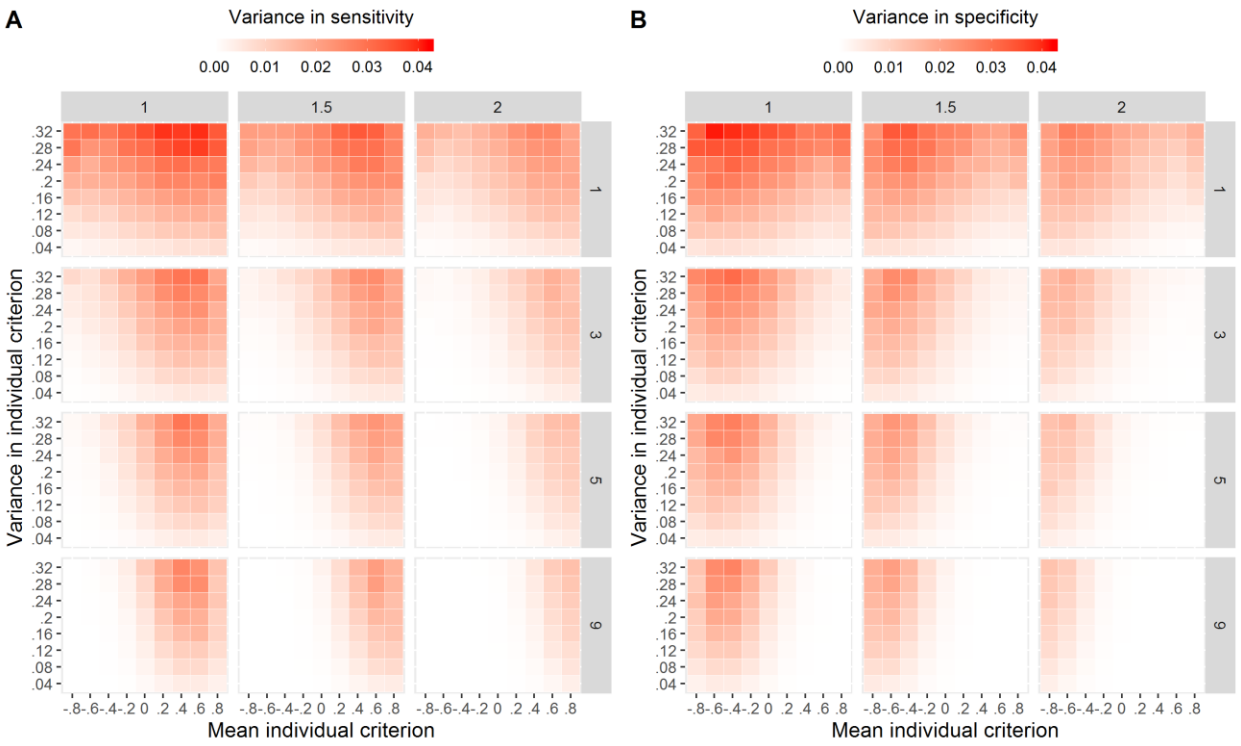

Fig. S2. Numerical simulations: Reduction in variation in sensitivity and specificity when pooling decision from larger groups. This figure corresponds to the simulations shown in Fig. 2C, D in the main text where mean and variance in individual response bias (criterion) are varied across populations. Next to a reduction in variance in response bias with increasing group size (Fig. 2D), we also observe that variance in (A) sensitivity and (B) specificity decreases with increasing group size. This is shown for four levels of group size (n = 1, 3, 5 and 9; rows), and three levels of discrimination ability (d' = 1, 1.5, 2; columns).

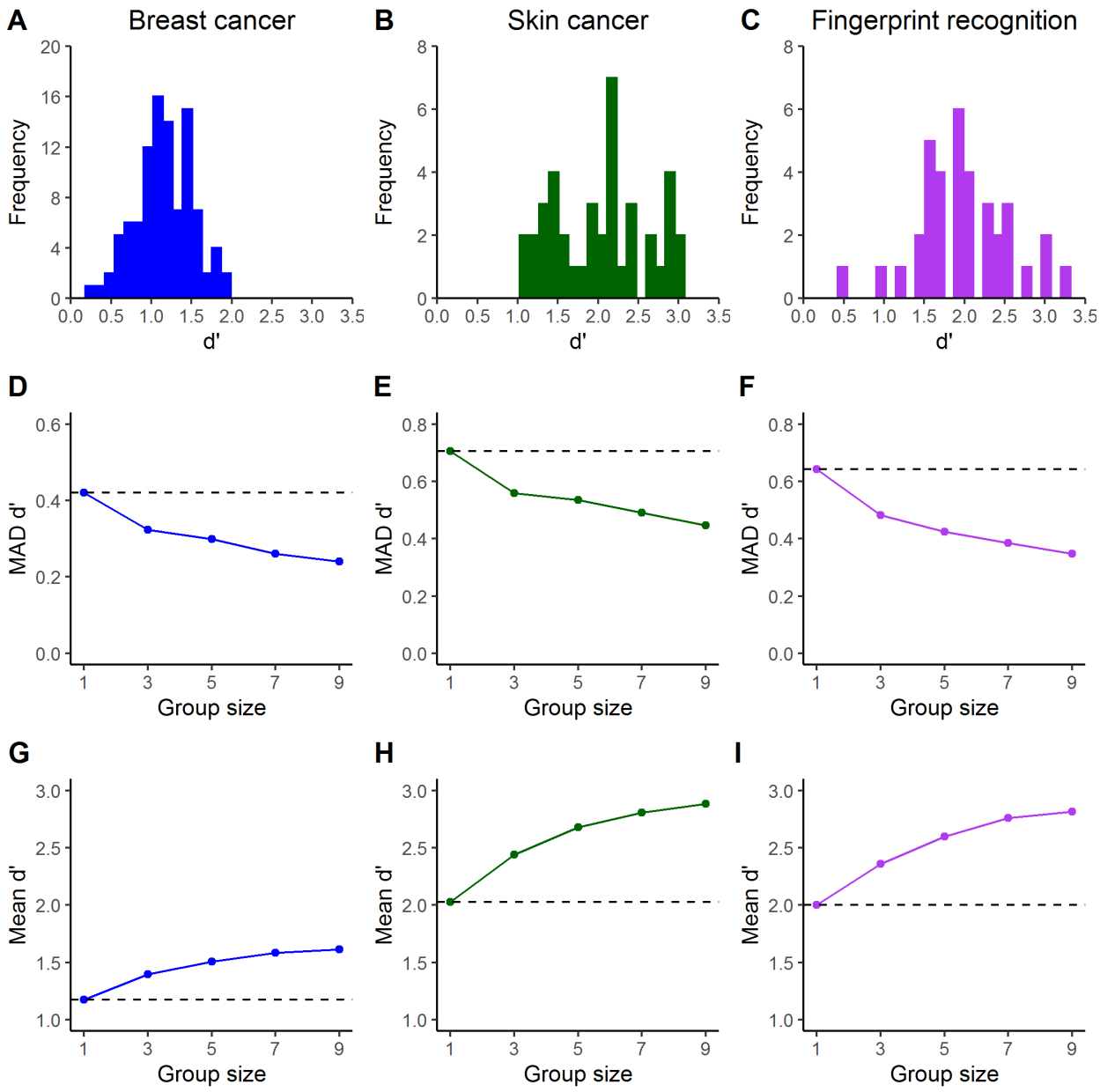

Fig. S3. Pooling independent decisions robustly reduces variation in *d'* in breast and skin cancer detection and fingerprint recognition. (A-C) Frequency distribution showing the distribution of *d'* levels of raters in the three datasets, showing that raters in all domains differ substantially in discrimination ability. (D-F) In all three datasets, compared to the baseline level of variation (mean absolute difference, MAD) in *d'* between individual decision makers (dashed lines), pooling independent decisions substantially decreases that level of variation in *d'* between decision-making agents (i.e. groups), as expected, these reductions increase with increasing group size. (G-I) In all cases, reductions in *d'* are accompanied by increased mean levels of *d'*.

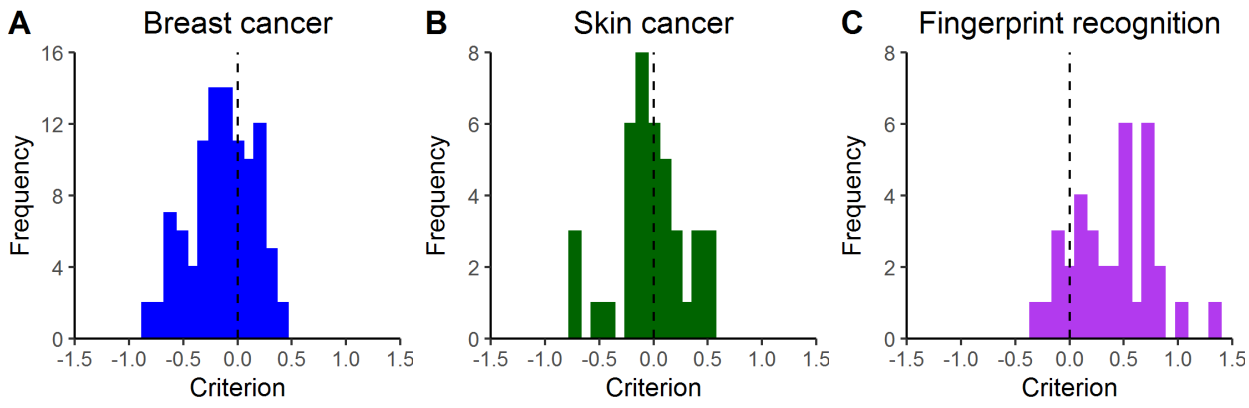

Fig. S4. Raters differ substantially in response bias. (A-C) Frequency distribution showing the distribution of criterion values of raters in the (A) breast cancer, (B) skin cancer, and (C) fingerprint recognition dataset, showing that raters in all these domains differ substantially in criterion values.

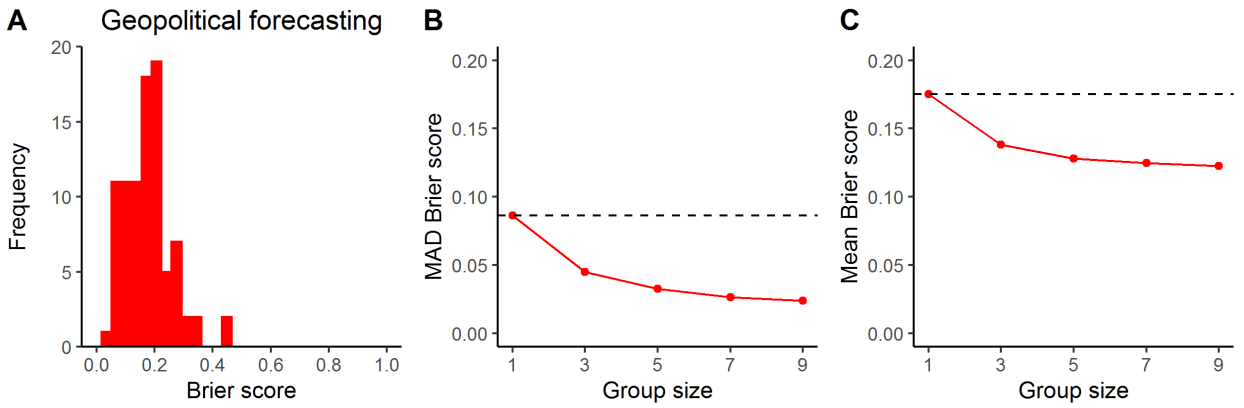

Fig. S5: Pooling independent decisions reduces variation in accuracy in geopolitical forecasting when using continuous forecast scores. (A) Frequency distribution showing the distribution of Brier scores of individual forecasters, showing substantial individual variation in Brier score (i.e. performance). (B) Compared to the baseline level of variation (mean absolute difference, MAD) in Brier score between individual decision makers (dashed line), pooling independent decisions substantially decreases that level of variation in Brier score between decision-making agents (i.e. groups), as expected, this reduction increases with increasing group size. (C) This reduction in variation is accompanied by a decrease in mean Brier score (i.e. more accurate forecasts).