Boosting medical diagnostics by pooling independent judgments

Ralf H. J. M. Kurvers\textsuperscript{a,b,1}, Stefan M. Herzog\textsuperscript{a}, Ralph Hertwig\textsuperscript{a}, Jens Krause\textsuperscript{b}, Patricia A. Carney\textsuperscript{c,d}, Andy Bogart\textsuperscript{e}, Giuseppe Argenziano\textsuperscript{f}, Iris Zalaudek\textsuperscript{g}, and Max Wolf\textsuperscript{b}

\textsuperscript{a}Center for Adaptive Rationality, Max Planck Institute for Human Development, 14195 Berlin, Germany; \textsuperscript{b}Department of Biology and Ecology of Fishes, Leibniz Institute of Freshwater Ecology and Inland Fisheries, 12587 Berlin, Germany; \textsuperscript{c}Department of Family Medicine, Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239; \textsuperscript{d}Department of Public Health & Preventive Medicine, Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239; \textsuperscript{e}Group Health Research Institute, Seattle, WA 98101; \textsuperscript{f}Department of Dermatology, Second University of Naples, 80131 Naples, Italy; and \textsuperscript{g}Department of Dermatology and Venerology, Medical University of Graz, 8036 Graz, Austria

Edited by Lee D. Ross, Stanford University, Stanford, CA, and approved June 7, 2016 (received for review February 2, 2016)

Collective intelligence refers to the ability of groups to outperform individual decision makers when solving complex cognitive problems. Despite its potential to revolutionize decision making in a wide range of domains, including medical, economic, and political decision making, at present, little is known about the conditions underlying collective intelligence in real-world contexts. We here focus on two key areas of medical diagnostics, breast and skin cancer detection. Using a simulation study that draws on large real-world datasets, involving more than 140 doctors making more than 20,000 diagnoses, we investigate when combining the independent judgments of multiple doctors outperforms the best doctor in a group. We find that similarity in diagnostic accuracy is a key condition for collective intelligence: Aggregating the independent judgments of doctors outperforms the best doctor in a group whenever the diagnostic accuracy of doctors is relatively similar, but not when doctors’ diagnostic accuracy differs too much. This intriguingly simple result is highly robust and holds across different group sizes, performance levels of the best doctor, and collective intelligence rules. The enabling role of similarity, in turn, is explained by its systematic effects on the number of correct and incorrect decisions of the best doctor that are overruled by the collective. By identifying a key factor underlying collective intelligence in two important real-world contexts, our findings pave the way for innovative and more effective approaches to complex real-world decision making, and to the scientific analyses of those approaches.

Significance

Collective intelligence is considered to be one of the most promising approaches to improve decision making. However, up to now, little is known about the conditions underlying the emergence of collective intelligence in real-world contexts. Focusing on two key areas of medical diagnostics (breast and skin cancer detection), we here show that similarity in doctors’ accuracy is a key factor underlying the emergence of collective intelligence in these contexts. This result paves the way for innovative and more effective approaches to decision making in medical diagnostics and beyond, and to the scientific analyses of those approaches.

Author contributions: R.H.J.M.K., S.M.H., R.H., J.K., and M.W. designed research; P.A.C., A.B., G.A., and I.Z. performed research; R.H.J.M.K., S.M.H., and M.W. analyzed data; and R.H.J.M.K. and M.W. wrote the paper with input from all other authors.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

\textsuperscript{1}To whom correspondence should be addressed. Email: kurvers@mpib-berlin.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1601827113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1601827113
dataset comprises 4,320 diagnoses and confidence estimates made by 40 dermatologists of 108 skin lesions (33), with a mean individual sensitivity ± SD = 0.833 ± 0.130 and specificity = 0.835 ± 0.070 (SI Appendix, Fig. S1). These datasets allowed us to investigate the performance of collective intelligence rules that are based on aggregating the independent judgments of multiple doctors, and how this performance depends on the similarity in doctors’ diagnostic accuracy (a discussion of approaches based on direct interactions between doctors is provided in Discussion).

Results
We investigated the performance of virtual groups of diagnosticians using either of two collective intelligence rules: the confidence rule (17, 20) and the majority rule (34, 35). For any particular group evaluating any particular case, the confidence rule adopts the judgment of the most confident diagnostican, whereas the majority rule adopts the judgment receiving the most support within that group. For any of our groups, we compared the performance of these two rules with the performance of the best diagnostican in that group in terms of (i) sensitivity, (ii) specificity, and (iii) Youden’s index (J). The last is a composite measure of accuracy that combines sensitivity and specificity (J = Sensitivity + Specificity − 1) (36, 37).

Pouring the independent judgments of multiple diagnosticians with the confidence or the majority rule can only promote collective intelligence when two conditions are fulfilled. First, the judgments of different diagnosticians must not be perfectly correlated with each other (if different diagnosticians give identical judgments on all cases, there is no scope for collective intelligence). Second, in case of the confidence rule, there has to be a positive correlation between confidence and accuracy levels. Initial analyses of our datasets showed that both conditions are fulfilled in both diagnostic contexts (Fig. 1).

We first considered groups of two diagnosticians using the confidence rule. In both diagnostic contexts, we found that as the difference in accuracy levels between two diagnosticians increases, their joint ability to outperform the better diagnostican decreases [breast cancer: estimate (est) ± SE = −1.03 ± 0.04, t = −24.9, P < 0.001, Fig. 2A; skin cancer: est ± SE = −0.55 ± 0.03, t = −20.1, P < 0.001, Fig. 2B]. When diagnosticians’ accuracy levels were relatively similar (|ΔJ| < 0.1), the confidence rule outperformed the better diagnostican (Fig. 2A and B). In contrast, for relatively dissimilar groups, the better diagnostican outperformed the confidence rule. This effect was largely independent of the accuracy level of the better diagnostican (Fig. 3A and B), the accuracy level of the poorer diagnostican (SI Appendix, Fig. S4A and B), and the average accuracy level within groups (SI Appendix, Fig. S5A and B).

When we analyzed the effects of similarity in accuracy on collective sensitivity and specificity, the same pattern emerged: In both diagnostic contexts, combining decisions using the confidence rule led to higher sensitivity and specificity (relative to the better individual), but only when the two diagnosticians’ accuracy levels were similar (SI Appendix, Fig. S6). Moreover, independent of similarity, the confidence rule consistently outperformed the average individual performance within the group (SI Appendix, Fig. S7). When considering groups of three and five diagnosticians using the confidence rule, we find that the above results generalize to these larger group sizes (SI Appendix, Fig. S8).

Similarly, when considering groups of three diagnosticians using the majority rule, we found that as the differences in accuracy levels across the three diagnosticians increase, the group’s joint ability to outperform the best diagnostican decreases in both diagnostic contexts [breast cancer: est ± SE = −1.26 ± 0.05, t = −27.7, P < 0.001, Fig. 2C; skin cancer: est ± SE = −0.68 ± 0.03, t = −23.3, P < 0.001, Fig. 2D]. As in the case of the confidence rule, the majority rule outperformed the best diagnostican in that group only when the three diagnosticians’ accuracy levels were relatively similar (|ΔJ| < 0.1). Again, this effect was largely independent of the performance of the best diagnostican (Fig. 3C and D), the performance of the poorest diagnostican (SI Appendix, Fig. S4C and D), and the average performance within groups (SI Appendix, Fig. S5C and D), and it held for both sensitivity and specificity (SI Appendix, Fig. S9). Moreover, independent of diagnostic similarity, the majority rule consistently outperformed the average individual performance within the group (SI Appendix, Fig. S10). When considering groups of five diagnosticians using the majority rule, we find that the above results generalize to this larger group size (SI Appendix, Fig. S11). SI Appendix, Fig. S12 provides a direct comparison of the confidence and the majority rule for different group sizes.

To further understand the mechanisms underlying the above findings, we developed simplified analytical models of the two most basic scenarios investigated above, namely, two diagnosticians using the confidence rule and three diagnosticians using the majority rule (model details are provided in SI Appendix). To illustrate, consider two diagnosticians using the confidence rule. From the point of view of the better (i.e., more accurate) diagnostican, employing the confidence rule has two effects: The poorer
dicts that as similarity decreases, the ability of the group to out-
and in line with our main findings above (Fig. 2), the model pre-
the poorer makes more incorrect judgments). As a consequence,
poorer diagnostician/the majority decreases; Fig. 4, green bars) and
judgments of the best diagnostician that were overruled by the
modeling analysis. In particular, we find that as similarity decreases,
the number of correct and incorrect judgments of the best diagnostician
holds for the situation where three diagnosticians use the majority
rule (the better diagnostician makes more correct judgments and
the poorer gets worse, thereby decreasing the positive effect (the
better diagnostician makes fewer incorrect judgments and the
poorer diagnostician may overrule incorrect judgments of the better di-
agnostician (positive effect), and the poorer diagnostician may
overrule correct judgments of the better diagnostician (negative effect).
Importantly, our model shows that the strength of both effects depends on the similarity in accuracy levels between the two
diagnosticians (SI Appendix). As similarity decreases (assuming constant average accuracy), the better diagnostician gets better and
the poorer gets worse, thereby decreasing the positive effect (the
better diagnostician makes fewer incorrect judgments and the
poorer makes fewer correct judgments) and increasing the negative
effect (the better diagnostician makes more correct judgments and
the poorer makes more incorrect judgments). As a consequence,
and in line with our main findings above (Fig. 2), the model predicts that as similarity decreases, the ability of the group to out-
perform its better member also decreases. An analogous trend holds for the situation where three diagnosticians use the majority
rule (SI Appendix).

Further analyses of our datasets showed that in both diagnostic
contexts and for both the confidence and the majority rule, the
number of correct and incorrect judgments of the best diagnostician
that are overruled is fully in line with the prediction from the above
modeling analysis. In particular, we find that as similarity decreases,
(i) the positive effect above decreases (i.e., the number of incorrect
judgments of the best diagnostician that were overruled by the
poorer diagnostician/the majority decreases; Fig. 4, green bars) and
(ii) the negative effect above increases (i.e., the number of correct
judgments of the best diagnostician that were overruled by the
poorer diagnostician/the majority increases; Fig. 4, red bars).
Moreover, while the positive effect outweighs the negative effect
for relatively high levels of similarity ($|\Delta J| < 0.1$), the reverse is true
for relatively low levels of similarity ($|\Delta J| > 0.2$), thereby explaining
why only relatively similar groups can successfully use the confi-
dence and majority rule to outperform their best member.

**Discussion**

Although collective intelligence has the potential to transform
decision making in a wide range of domains, little is known about
the conditions that underlie its emergence in real-world contexts.
Drawing on large real-world datasets, involving more than 140
doctors performing more than 20,000 diagnoses, we have identi-
ified similarity in decision accuracy as a key factor underlying the emergence of collective intelligence in breast and skin cancer di-
gnostics. In particular, we have found that when a group of di-
agnosticians is characterized by relatively similar accuracy levels,
combining their independent judgments improves decision accu-
tracy relative to the best diagnostician within that group. In contrast,
when accuracy levels become too disparate, combining independent judgments leads to poorer diagnostic outcomes relative to those
diagnostic outcomes achieved by the best diagnostician. This re-
result is highly robust and holds across different performance levels
of the best diagnostician, different group sizes, and different col-
lective intelligence rules (confidence rule and majority rule).

To reap the benefits associated with collective intelligence, we
need to know which characteristics of decision makers and de-
cision contexts favor the emergence of collective intelligence and
which decision-making rules allow this potential to be harnessed.
We have here provided answers to both questions in the domain

![Fig. 2. Performance difference between the confidence/majority rule and the best diagnostician in a group as a function of the difference in accuracy levels (i.e., $|\Delta J|$) between diagnosticians. Results are shown for groups of two diagnosticians using the confidence rule (A and B) and for groups of three diagnosticians using the majority rule (C and D). Each dot represents a unique combination of two (or three) diagnosticians. Values above 0 indicate that the confidence/majority rule outperformed the best individual in the group. Values below 0 indicate that the best individual outperformed the confidence/majority rule. Red lines are linear regression lines. In both breast cancer (A and C) and skin cancer (B and D) diagnostics, the confidence/majority rule outperformed the best individual only when the diagnosticians’ accuracy levels were relatively similar ($|\Delta J| < 0.1$).]

![Fig. 3. Performance difference between the confidence/majority rule and the best diagnostician in a group as a function of the difference in accuracy levels between diagnosticians and the accuracy level of the best diagnostician. Shown are results for groups of two diagnosticians using the confidence rule (A and B) and for three diagnosticians using the majority rule (C and D). Red areas indicate that the confidence/majority rule outperformed the best diagnostician within that group, white areas indicate no performance difference, and gray and black areas indicate that the best diagnostician outperformed the confidence/majority rule. Shown are averaged values based on (maximally 1,000) randomly drawn unique groups. The confidence/majority rule outperformed the best diagnostician only when the diagnosticians’ accuracy levels were relatively similar (i.e., left part of the heat plots). This effect was independent of the accuracy level of the best diagnostician.]

Kurvers et al.
Box plots show medians and interquartile ranges. As predicted from our drawn unique groups, using either of the two collective intelligence rules. three onward, the majority rule tends to outperform the confidence employed to outperform the best diagnostician. For a group size of member. At a group size of two, the confidence rule can be suggest that groups of diagnosticians with similar accuracy levels of mammograms, a standard practice in Europe (38). Our analyses or the majority (\( j \)) within a group was overruled by the more confident diagnostician (diag.) within a group was overruled by the more confident diagnostician (A and B) or the majority (C and D). Red box plots correspond to the number of cases where a correct decision of the best diagnostician within a group was overruled by the more confident diagnostician (A and B) or the majority (C and D). Shown are averaged values based on (maximally 1,000) randomly drawn unique groups, using either of the two collective intelligence rules. Box plots show medians and interquartile ranges. As predicted from our modeling analysis (SI Appendix), with decreasing similarity in accuracy levels (i.e., higher \(| Δ j |\)), the number of incorrect decisions by the best individual that were overruled decreased and the number of correct decisions by the best individual that were overruled increased.

Future studies should address at least three issues. First, we have focused on combining independent diagnostic judgments, thus not investigating situations in which diagnosticians directly communicate with each other. Therefore, one open question is the extent to which our findings generalize to face-to-face interactions and discussions within medical teams. Previous work in nonmedical contexts has shown that similarity in accuracy is a prerequisite for collective intelligence to arise during group discussions (15), suggesting that it may also matter in interacting medical teams. Second and more generally, it remains unknown how these two collective intelligence mechanisms (aggregation of independent judgments versus group discussions) compare in medical diagnostics (19). Group discussions are known to be a double-edged sword (39). Phenomena such as group think, interpersonal competition, social loafing, and obedience to authority can compromise group accuracy (40–42), yet groups are known to outperform individuals across a range of tasks (7, 43). It will thus be important to compare the relative gains (or declines) in accuracy that these mechanisms afford across medical diagnostic contexts. Third, improving decision accuracy is of prime importance across a wide range of contexts (e.g., economic decision making, political decision making). Future work should investigate whether and to what extent similarity in decision accuracy is a key enabling factor of collective intelligence in these contexts.

Materials and Methods

Our analyses are based on the two previously published datasets outlined below.

Breast Cancer Dataset. The breast cancer dataset comprises 16,813 interpretations of 182 mammograms made by 101 radiologists (mean number of mammograms evaluated per radiologist = 166, range: 161–173) and is one of the largest mammography datasets available (32). Mammograms included in the test set were randomly selected from screening examinations performed on women aged 40–69 y between 2000 and 2003 from US mammography registries affiliated with the Breast Cancer Surveillance Consortium (BCSC; Carolina Mammography Registry, New Hampshire Mammography Network, New Mexico Mammography Project, Vermont Breast Cancer Surveillance System, and Group Health Cooperative in western Washington). Radiologists who interpreted mammograms at facilities affiliated with these registries between January 2005 and December 2006 were invited to participate in this study, as well as 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 (32). Women who were diagnosed with cancer within 12 mo of the mammograms were classified as cancer patients (n = 51). Women who remained cancer-free for a period of 2 y were classified to be noncancer patients (n = 131; i.e., 28% prevalence of cancer).

Radiologists viewed the digitized images on a computer (home computer, office computer, or laptop provided as part of the original study). The computers were required to meet all viewing requirements of clinical practice, including a large screen and high-resolution graphics (\( \geq 1,280 \times 1,024 \) pixels and a 1280MB video-card with 32-bit color). Radiologists saw two images at the same time (i.e., the left and right breasts) and were able to alternate quickly (\( \leq 1 \) 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 workup).

Radiologists evaluated the cases in two stages. For stage 1, four test sets were created, with each containing 109 cases (32). Radiologists were randomly assigned to one of the four test sets. For 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 part 1. To avoid having the same radiologist evaluating a case twice, we excluded all cases from part 2 that had already been viewed by that radiologist in part 1. This procedure resulted in a total of 161 unique cases for radiologists in test sets 1 (n = 25 radiologists) and 2 (n = 30 radiologists) and 173 unique cases for radiologists in test sets 3 (n = 26 radiologists) and 4 (n = 20 radiologists), resulting in a total of 16,813 unique readings. Between the two stages, radiologists were randomly assigned to one of three intervention treatments. Because there were no strong treatment differences (44), we pooled the data from stages 1 and 2. For all group simulation analyses, radiologists were always grouped within the four test sets (because radiologists in the same test set had evaluated the same images).

In our analysis, we treated the recommendation that a woman should be recalled for further examination as a positive test result. After providing each final diagnosis, radiologists rated their confidence in it on a five-point scale.

Skin Cancer Dataset. The skin cancer dataset comprises 4,320 diagnoses by 40 dermatologists of 108 skin lesions and was collected as part of a consensus meeting via the internet, called the Consensus Net Meeting on Dermoscopy (33). 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, Aus-
tra利亚); and Skin and Cancer Associates (Plantation, FL). The lesions were se-
lected based on the photographic quality of the clinical and dermoscopic images available. The goal of the study was to diagnose whether 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, nonmelanoma: n = 81; i.e., 25% prevalence).

All participating dermatologists had at least 5 years of experience in dermoscopy practice, teaching, and research. They first underwent 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, the dermatologist was asked to classify it as either melanoma or a benign melanocytic lesion, using four different algorithms. Here, we focus on the diagnostic algorithm with the highest diagnostic accuracy which is also the one most widely used for melanoma detection: pattern analysis (33). It uses a set of global (textured patterns covering most of the lesion) and local features (representing characteristics that appear in part of the lesion) to differentiate between melanomas and benign melanocytic lesions.

We next classified each skin lesion to classify a lesion as melanoma as a positive test result. After providing each final diagnosis, dermatologists rated their con-


difidence in it on a four-point scale.

Ethics Statement. The breast cancer data were assembled at the BCSC Statistical Coordinating Center (SCC) in Seattle and analyzed at the Leibniz Institute of Freshwater Ecology and Inland Fisheries in Berlin (IGB), Germany. Each registry, the SCC, and the IGB, received institutional review board approval for active and passive consent processes or were granted a waiver of consent to enroll par-
ticipants, 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 breastscreening.cancer.gov/work/proposal_data.html.

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 has been included in Dataset S1.

Collective Intelligence Rules. Both datasets include the judgments of experts who independently evaluated the same cases and rated their confidence in each diagnosis. We created virtual groups of diagnosticians who evaluated the cases “together” using two collective intelligence rules: the confidence rule (17, 20) and the majority rule (34, 35).

Confidence rule. Separately for both datasets, we created virtual groups (for group sizes of two, three, and five diagnosticians). For each group size, we set an upper limit of 1,000 randomly drawn unique groups. The confidence rule stipulates that the diagnosis of the most confident diagnostican in the group is adopted, irrespective of the confidence associated with those decisions. We classified each case as “cancerous” or “noncancerous” depending on whether the two diagnoses received more votes among the group members. We then evaluated the performance of the majority rule for each group in terms of (i) sensitivity, (ii) specificity, and (iii) Youden’s index (J). For a group size of two, this measure is simply the abso-
lute difference in J between the two group members. For a group size of three or more, this measure is the expected absolute difference in J between two randomly chosen group members. We analyzed the effect of similarity in accuracy on a group’s ability to outperform its best individual(s) using general linear models in R (version 3.2.2). Significance levels were de-

rived from the t values and associated P values.

ACKNOWLEDGMENTS. We thank the editor and three anonymous referees for numerous helpful suggestions on a previous version of our manuscript. We thank Susannah Goss, Tim Pleskac, Juliane Kämmer, Aleksandra Litvinova, and members of the Center for Adaptive Rationality for helpful comments on earlier versions of the manuscript. We thank Jose Cayere, Amy Buzby, and the American College of Radiology for their technical assistance in de-
v eloping 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 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-

08-270-01, SIRSG-09-270-01, SIRSG-09-271, SIRSG-09-270-04); by the Breast Cancer Stamp Fund; and by the National Cancer Institute Breast Can-
cer 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 through-

out the United States. A full description of these sources is provided at www.breastscreening.cancer.gov/work/acknowledgement.html.
Supporting Information

**Boosting medical diagnostics by pooling independent judgments**

Ralf H.J.M. Kurvers, Stefan M. Herzog, Ralph Hertwig, Jens Krause, Patricia A. Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek & Max Wolf
Fig. S1: Histograms of average individual performance.

The frequency of average individual (A, B) sensitivity, (C, D) specificity and (E, F) Youden’s index of the (A, C, E) radiologists (n = 101) and (B, D, F) dermatologists (n = 40).
Fig. S2: Performance of the confidence/majority rule relative to the best diagnostican in that
group as a function of the independence of judgments.

(A-D) Each dot represents a unique combination of (A, B) two or (C, D) three diagnosticians. Values
above zero indicate that the confidence/majority rule outperformed the best individual in that group.
Values below zero indicate that the best diagnostican outperformed the confidence/majority rule. Red
lines are linear regression lines. With increasing Cohen’s kappa (i.e. lower independence of
judgments), the ability of groups to outperform its best member decreases. Only groups with
diagnosticians of similar ability in terms of Youden’s index (i.e. ΔJ < 0.1) are shown since this is the
region where collective intelligence arises.
Fig. S3: Relationship between the reported confidence level and the sensitivity/specificity.

There was a positive relationship between confidence and sensitivity/specificity for the (i) best- (ii) midlevel-, and (iii) poorest performing diagnosticians (based on the Youden’s index). For a given confidence level, the performance of the best diagnosticians was generally higher than the performance of the middle diagnosticians which, in turn, was generally higher than the performance of the poorest diagnosticians.
Fig. S4: Difference in performance between the confidence/majority rule and the best diagnostician in the group as a function of the performance of the poorest diagnostician. Difference in performance between the collective intelligence rule and the best group member as a function of the difference in accuracy (Youden’s index $J$) between group members (x-axis) and the accuracy of the worst group member (y-axis). Shown are results for groups of (A, B) two diagnosticians using the confidence rule, and (C, D) three diagnosticians using the majority rule. Red areas indicate that the confidence/majority rule outperformed the best diagnostician; white areas indicate no difference; grey and black areas indicate that the best diagnostician outperformed the confidence/majority rule. The confidence/majority rule outperformed the best diagnostician only when the diagnosticians’ accuracy levels were similar (i.e., left part of the heat plots). This effect was present both in groups in which the worst diagnostician performed relatively well and in groups in which he/she performed poorly.
Fig. S5: Difference in performance between the confidence/majority rule and the best diagnosticians in the group as a function of average individual performance.

Difference in performance between the collective intelligence rule and the best group member as a function of the difference in accuracy (Youden’s index, $J$) between group members (x-axis) and the average individual performance of group members (y-axis). Shown are results for groups of (A, B) two diagnosticians using the confidence rule and (C, D) three diagnosticians using the majority rule. Red areas indicate that the confidence/majority rule outperformed the best diagnostician; white areas indicate no difference; grey and black areas indicate that the best diagnostician outperformed the confidence/majority rule. The confidence/majority rule outperformed the best diagnostician only when the diagnosticians’ accuracy levels were similar (i.e., left part of the heat plots). This effect was present both in groups in which the average individual ability of diagnosticians was high and in groups in which average individual ability of diagnosticians was low.
Fig. S6: Sensitivity and specificity of the confidence rule in groups of two diagnosticians relative to the best diagnostician in that group.

(A-D) Each dot represents a unique combination of two diagnosticians. Values above zero indicate that the confidence rule outperformed the best diagnostician in that group; values below zero indicate that the best diagnostician outperformed the confidence rule. Red lines are linear regression lines. In both diagnostic contexts, the confidence rule achieved higher (A, B) sensitivity and (C, D) specificity than the best diagnostician only when the diagnosticians’ accuracy levels were similar.
Fig. S7: Performance of the confidence rule in groups of two diagnosticians relative to average individual performance in that group.

(A, B) Each dot represents a unique combination of two diagnosticians. Values above zero indicate that the confidence rule outperformed the average individual in that group; values below zero indicate that the average individual outperformed the confidence rule. In both (A) breast and (B) skin cancer diagnostics, the confidence rule generally outperformed average individual performance (proportion of groups in which the confidence rule outperformed the average performance: breast cancer: 0.84; skin cancer: 0.80).
Fig. S8: Performance of the confidence rule in groups of three and five diagnosticians relative to the best diagnostican in that group.

(A-D) Each dot represents a unique combination of (A, B) three or (C, D) five diagnosticians. Values above zero indicate that the confidence rule outperformed the best diagnostican in that group. Values below zero indicate that the best diagnostican outperformed the confidence rule. Red lines are linear regression lines. In both diagnostic contexts and for both group sizes, the confidence rule outperformed the best diagnostican only when diagnosticians’ accuracy levels were relatively similar.
Fig. S9: Sensitivity and specificity of the majority rule in groups of three diagnosticians relative to the best diagnostican in that group.

(A-D) Each dot represents a unique combination of three diagnosticians. Values above zero indicate that the majority rule outperformed the best diagnostican in that group. Values below zero indicate that the best diagnostican outperformed the majority rule. Red lines are linear regression lines. In both diagnostic contexts, the majority rule achieved higher (A, B) sensitivity and (C, D) specificity than the best diagnostican only when diagnosticians’ accuracy levels were relatively similar.
**Fig. S10**: Performance of the majority rule in groups of three diagnosticians relative to average individual performance in that group.

(A, B) Each dot represents a unique combination of three diagnosticians. Values above zero indicate that the majority rule outperformed the average individual in that group. Values below zero indicate that the average individual outperformed the majority rule. In both (A) breast and (B) skin cancer diagnostics, the majority rule generally outperformed average individual performance (proportion of groups in which the majority rule outperformed the average performance: breast cancer: 0.98; skin cancer: 0.99).
Fig. S11: Performance of the majority rule in groups of five diagnosticians relative to the best diagnostian in that group.

(A, B) Each dot represents a unique combination of five diagnosticians. Values above zero indicate that the majority rule outperformed the best individual in that group. Values below zero indicate that the best diagnostian outperformed the majority rule. Red lines are linear regression lines. In both diagnostic contexts, the majority rule outperformed the best diagnostian only when diagnosticians’ accuracy levels were relatively similar.
Fig. S12: Comparing the performance of the confidence and majority rule across different group sizes.

Difference in performance between the confidence/majority rule and the best group member, grouped for three categories of similarity in performance ($\Delta J$), for group size (A, B) two, (C, D) three and (E, F) five. For all group sizes, collective intelligence rules should only be used when diagnosticians have similar performance level (i.e., $\Delta J<0.1$). If so, then at group size two the confidence rule performs best, at group size three and beyond the majority rule performs best.
PART II: MODELLING ANALYSIS

In order to further understand the mechanisms underlying our results and to investigate the generality of our findings, we developed simplified analytical models of the basic scenarios investigated in our data sets. In particular, we developed models for the scenarios where two diagnosticians employ the confidence rule (Model 1) and three diagnosticians employ the majority rule (Model 2). As can be seen in the following, the results of the analytical models are fully in line with the findings of the empirical analysis.

MODEL 1: Two diagnosticians employing the confidence rule

We consider two diagnosticians with probabilities of making a correct decision $p_1$ and $p_2$, respectively. Without loss of generality, we can assume that diagnostician 1 is the more accurate of the two diagnosticians, that is $p_1 > p_2$.

We are interested in the following question. Assuming the dyad adopts the confidence rule, how does the degree of similarity in the two diagnosticians’ accuracy levels affect the dyad’s ability to outperform the better diagnostician?

In order to study this question, we make the simplifying assumption that, for any of the two diagnosticians, the probability of being correct, conditional on the other diagnostician being correct, corresponds to the unconditional probability of being correct, that is:

$$p_i \bigg|_{\text{rater } j \text{ correct}} = p_i, \quad i, j = 1, 2 \text{ and } i \neq j. \quad (1)$$

The better diagnostician is outperformed whenever the probability that either (i) both diagnosticians are correct or (ii) one of the diagnosticians is correct and simultaneously has
the higher confidence score is higher than the probability that the better diagnostician is correct. That is, the better diagnostician is outperformed whenever:

\[ p_1 \cdot p_2 + (p_1 \cdot (1-p_2) + (1-p_1) \cdot p_2) \cdot \alpha > p_1, \] (2)

where \( \alpha \) corresponds to the probability that – in the case of a disagreement – the individual with the higher confidence score is correct.

This condition can be simplified to:

\[ (1-p_1) \cdot p_2 \cdot \alpha - p_1 \cdot (1-p_2) \cdot (1-\alpha) > 0. \] (3)

Since we are interested how the similarity in accuracy between diagnostician 1 and 2 affects the ability of the dyad to outperform the better diagnostician (while keeping the average ability in the dyad constant), we now rewrite \( p_1 \) and \( p_2 \) as:

\[ p_1 = \bar{p} + \delta, \]
\[ p_2 = \bar{p} - \delta, \] (4)

where \( \bar{p} \) corresponds to the average ability of the two diagnosticians (i.e. \( \frac{p_1 + p_2}{2} \)) and \( \delta \) is a measure of similarity between the two diagnostician (i.e. as \( \delta \) increases, the similarity between the two diagnosticians decreases). We now substitute (4) into (3):

\[ (1-(\bar{p}+\delta)) \cdot (\bar{p}-\delta) \cdot \alpha - (\bar{p}+\delta) \cdot (1-(\bar{p}-\delta)) \cdot (1-\alpha) > 0. \] (5)
Results Model 1

**Result 1.1:** For diagnosticians with identical performance, i.e. $\delta = 0$, the confidence rule allows dyads to outperform any of the diagnosticians whenever the probability $\alpha$ with which – in case of disagreement – the individual with the higher confidence is correct is larger than 0.5, that is whenever:

\[ \alpha > \frac{1}{2}. \]  

(6)

**Proof.** Substitute $\delta = 0$ in (5) and solve for $\alpha$.

**Result 1.2:** As the similarity between the two diagnosticians decreases (i.e. $\delta$ increases), the probability with which the dyad outperforms the better diagnostician decreases. More technically, the derivative of the left hand side of (5) with respect to $\delta$ is strictly negative. We note that, because of (4), this result applies to groups of differing similarity but identical average ability. Result 1.2 is illustrated in Fig. S13 below.

**Proof.** Taking the derivative of the left hand side of (5) with respect to $\delta$ results in

\[-(\bar{p} - \delta) \cdot \alpha - (1 - (\bar{p} + \delta)) \cdot \alpha - (1 - (\bar{p} - \delta)) \cdot (1 - \alpha) - (\bar{p} + \delta) \cdot (1 - \alpha), \]  

(7)

which, because $0 \leq \bar{p} + \delta \leq 1$, $0 \leq \bar{p} - \delta \leq 1$ and $0 \leq \alpha \leq 1$, is strictly negative. This establishes Result 1.2.

Inspecting the two main terms on the left hand side of (5), we can also get a good intuition for the mechanism underlying Result 1.2. As the similarity between the two diagnosticians decreases (i.e. $\delta$ increases) two regularities simultaneously hold:
(i) The probability that the poorer diagnostician overrules an incorrect decision by the better diagnostician decreases because the better makes fewer incorrect decisions and the poorer makes fewer correct decisions.

(ii) The probability that the poorer diagnostician overrules a correct decision by the better diagnostician increases because the better makes more correct decisions and the poorer makes more incorrect decisions.

Result 1.3: Consider scenarios in which the probability that, in case of a disagreement, the individual with the higher confidence score is correct is larger than 0.5 and smaller than 1, that is, $0.5 < \alpha < 1$. In these scenarios, for low levels of similarity (i.e., high $\delta$), the dyad performs worse than the better diagnostician; conversely, for high levels of similarity (i.e. low $\delta$), the dyad outperforms the better diagnostician. More specifically, for any given average performance level $\bar{p}$ of the two diagnosticians in the dyad, and any level $\alpha$ ($0.5 < \alpha < 1$), there exists a threshold level of similarity $\delta^*$ with the feature that dyads with a lower similarity (i.e. $\delta > \delta^*$) are outperformed by the better individual while dyads with a higher similarity (i.e. $\delta < \delta^*$) outperform the better individual. Result 1.3 is illustrated in Fig. S13 below.

Proof. From Result 1.1 we know that – whenever $\alpha > 0.5$ – the left hand side of (5) is positive for $\delta = 0$. Moreover, from Result 1.2 we know that the left hand side of (5) strictly decreases in $\delta$. To establish Result 1.3 it is thus sufficient to show that for any particular combination of $\bar{p}$ and $\alpha$ ($0.5 < \alpha < 1$), there exists a $\delta$ that turns the left hand side of (5) negative.

For this purpose, let us assume that

$$\delta = 1 - \bar{p}. \quad (8)$$
Note that this restricts our analysis to scenarios with $\bar{p} > 0.5$. Substituting (8) into (5) and rearranging the left hand side leaves us with

$$-(2 - 2 \cdot \bar{p}) \cdot (1 - \alpha),$$

which, because $0 < \bar{p} < 1$ and $0.5 < \alpha < 1$, is always negative. This establishes Result 1.3.
Fig. S13. As the similarity between two diagnosticians decreases (i.e. $\delta$ increases), the probability with which the dyad (when adopting the confidence rule) outperforms the better diagnostican decreases. (A) Illustrates this effect for three different levels of $\alpha$ (i.e. probabilities that – in case of a disagreement – the individual with the higher confidence score is correct) and a dyad with an average ability $\bar{p} = 0.7$. (B) Illustrates this effect for dyads with three different average abilities $\bar{p}$ and $\alpha = 0.75$. 
MODEL 2: Three diagnosticians employing the majority rule

We consider three diagnosticians with probabilities of making a correct decision \( p_1, p_2 \) and \( p_3 \), respectively. Without loss of generality we can assume that diagnosticians 1 is the most accurate of the three diagnosticians, that is \( p_1 > p_2 \) and \( p_1 > p_3 \).

We are interested in the following question. How does the degree of similarity in the three diagnosticians’ accuracy levels affect the ability of that group, assuming they adopt the majority rule, to outperform the best diagnosticians?

In order to analyze this question, we make two simplifying assumptions:

(i) For any of the three diagnosticians, the probability of being correct, conditional on any other diagnostician being correct corresponds to the unconditional probability of being correct, that is:

\[
\begin{align*}
\left. p_i \right| \text{rater } j \text{ correct} &= p_i, \quad i, j = 1, 2, 3 \text{ and } i \neq j. \tag{10}
\end{align*}
\]

(ii) The two poorer diagnosticians are identical in performance, that is \( p_2 = p_3 \).

Since we are interested how the similarity between the diagnosticians affects the ability of the three diagnosticians to outperform the best diagnosticians (while keeping the average ability in the group constant), we now rewrite \( p_1, p_2, \) and \( p_3 \) as:

\[
\begin{align*}
p_1 &= \bar{p} + \delta, \\
p_2 &= p_3 = \bar{p} - \frac{1}{2} \delta, \tag{11}
\end{align*}
\]

where \( \bar{p} \) corresponds to the average ability of three diagnosticians (i.e. \( \frac{p_1 + p_2 + p_3}{3} \)) and \( \delta \) is a measure of similarity between the diagnosticians (i.e. as \( \delta \) increases, the similarity between the best diagnosticians and the poorer diagnosticians decreases).
Under the majority rule, the best diagnostician is outperformed whenever the probability that at least two diagnosticians are correct is higher than the probability that the best diagnostician is correct, that is the best diagnostician is outperformed whenever:

\[ p_1 \cdot p_2 \cdot p_3 + p_1 \cdot (1 - p_3) + p_1 \cdot (1 - p_2) \cdot p_3 + (1 - p_1) \cdot p_2 \cdot p_3 > p_1. \]  

(12)

This condition can be simplified to:

\[ \frac{(1 - p_1) \cdot p_2 \cdot p_3}{\text{probability that an incorrect decision by the best rater is improved by the two poorer raters}} - \frac{p_1 \cdot (1 - p_2) \cdot (1 - p_3)}{\text{probability that a correct decision by the best rater is worsened by the two poorer raters}} > 0. \]  

(13)

We now substitute (11) into (13):

\[ \frac{(1 - (\bar{p} + \delta)) \cdot (\bar{p} - \frac{1}{2} \cdot \delta)^2}{\text{probability that an incorrect decision by the best rater is improved by the two poorer raters}} - \frac{(\bar{p} + \delta) \cdot \left(1 - \left(\bar{p} - \frac{1}{2} \cdot \delta\right)\right)^2}{\text{probability that a correct decision by the best rater is worsened by the two poorer raters}} > 0 \]  

(14)

Results Model 2

Result 2.1: In case of diagnosticians with identical performance, i.e. \( \delta = 0 \), condition (14) reduces to the Condorcet condition, that is, it is fulfilled whenever \( \bar{p} > 0.5 \).

Proof. Substitute \( \delta = 0 \) in (14) and solve for \( p \).

Result 2.2: As the similarity between diagnosticians decreases (i.e. \( \delta \) increases), the probability with which the group outperforms the best diagnostician within that group decreases. More technically, the derivative of the left hand side of (14) with respect to \( \delta \) is strictly negative, for all \( \bar{p} \) and \( \delta \). Result 2.2 is illustrated in Fig. S14 below.

Proof. Taking the derivative of the left hand side of (14) with respect to \( \delta \) results in:
\[-\left(\bar{p} - \frac{1}{2} \cdot \delta\right)^2 - (1 - (\bar{p} + \delta)) \cdot \left(\bar{p} - \frac{1}{2} \cdot \delta\right) - (1 - \left(\bar{p} - \frac{1}{2} \cdot \delta\right)) \cdot (1 - (\bar{p} + \delta)) \right] , \tag{15}\]

which – since \(0 \leq \bar{p} + \delta \leq 1\) and \(0 \leq \bar{p} - \frac{1}{2} \cdot \delta \leq 1\) – is strictly negative. This establishes Result 2.2.

Inspecting the two main terms on the left hand side of (14), we can also get a good intuition for this result. As the similarity between diagnosticians decreases (i.e. \(\delta\) increases), the following two regularities simultaneously hold:

(i) The probability that the two poorer diagnosticians overrule an incorrect decision by the best diagnostician decreases because the best makes fewer incorrect decisions and the poorer make fewer correct decisions.

(ii) The probability that the two poorer diagnosticians overrule a correct decision by the best diagnostician increases because the best makes more correct decisions and the poorer make more incorrect decisions.

**Result 2.3:** Consider scenarios with \(\bar{p} > 0.5\). For low levels of similarity (i.e. high \(\delta\)), the group performs worse than the best diagnostician; conversely, for high levels of similarity (i.e. low \(\delta\)), the group outperforms the best diagnostician. More specifically, for any given average performance level \(\bar{p}\) in the group, there exists a threshold level of similarity \(\delta^*\) with the feature that groups with a lower similarity (i.e., \(\delta > \delta^*\)) are outperformed by the best individual while groups with a higher similarity (i.e., \(\delta < \delta^*\)) outperform the best individual. Result 2.2 is illustrated in Fig. S14 below.

**Proof.** From Result 2.1 we know that the left hand side of (14) is positive for \(\bar{p} > 0.5\). From Result 2.2 we know that the left hand side of (14) strictly decreases in \(\delta\). To establish Result
2.3 it is thus sufficient to show that for any given $\bar{p}$ there exists one $\delta$ which turns equation (14) negative.

For this purpose, let us assume that

$$\delta = 1 - \bar{p}. \quad (16)$$

Substituting (16) into (14) and rearranging the left hand side leaves us with

$$-\left(\frac{3}{2}(1 - \bar{p})\right)^2, \quad (17)$$

which is always negative. This establishes Result 2.3.
**Fig. S14.** As the similarity between the three diagnosticians decreases (i.e. $\delta$ increases), the probability with which these three diagnosticians – when adopting the majority rule – outperform the best diagnostian decreases. This effect is illustrated for three different average abilities $\bar{p}$ of the three diagnosticians.