PLOS ONE

RESEARCH ARTICLE

Combining independent decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging

Ralf H. J. M. Kurvers[1]*, Annemarie de Zoete[2], Shelby L. Bachman[1], Paul R. Algra[3], Raymond Ostelo[2]

1 Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee, Berlin, Germany, **2** Department of Health Sciences, Amsterdam Public Health research institute, Vrije Universiteit, Amsterdam, the Netherlands, **3** Department of Radiology, Medical Centre Alkmaar, Alkmaar, the Netherlands

* kurvers@mpib-berlin.mpg.de

Citation: Kurvers RHJM, de Zoete A, Bachman SL, Algra PR, Ostelo R (2018) Combining independent decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging. PLoS ONE 13(4): e0194128. https://doi.org/10.1371/journal.pone.0194128

Editor: Gonzalo G. de Polavieja, Fundacao Champalimaud, PORTUGAL

Received: May 23, 2017

Accepted: February 26, 2018

Published: April 3, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: R. Ostelo acknowledges funding of the Netherlands Organisation for Health Research and Development (ZonMW), the Netherlands Organisation for Scientific Research (NWO), the Scientific College of Physiotherapy (WCF) and EUROSPINE. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Abstract

Diagnosing the causes of low back pain is a challenging task, prone to errors. A novel approach to increase diagnostic accuracy in medical decision making is collective intelligence, which refers to the ability of groups to outperform individual decision makers in solving problems. We investigated whether combining the independent ratings of chiropractors, chiropractic radiologists and medical radiologists can improve diagnostic accuracy when interpreting diagnostic images of the lumbosacral spine. Evaluations were obtained from two previously published studies: study 1 consisted of 13 raters independently rating 300 lumbosacral radiographs; study 2 consisted of 14 raters independently rating 100 lumbosacral magnetic resonance images. In both studies, raters evaluated the presence of "abnormalities", which are indicators of a serious health risk and warrant immediate further examination. We combined independent decisions of raters using a majority rule which takes as final diagnosis the decision of the majority of the group. We compared the performance of the majority rule to the performance of single raters. Our results show that with increasing group size (i.e., increasing the number of independent decisions) both sensitivity and specificity increased in both data-sets, with groups consistently outperforming single raters. These results were found for radiographs and MR image reading alike. Our findings suggest that combining independent ratings can improve the accuracy of lumbosacral diagnostic image reading.

Introduction

Low back pain is a major health problem in the industrialized world. In the United States it is, for example, the second most common reason for visiting health care workers, and approximately 1/3 of adults in the US reports low back pain during the last 3 months [1–5]. Accurately diagnosing the causes of low back pain is, however, known to be particularly challenging [6–8]. A widely used method to aid health care providers in diagnosing the causes of low back

Competing interests: The authors have declared that no competing interests exist.

pain is the use of lumbar and lumbosacral spine imaging (i.e., radiography, computerized tomography (CT) and magnetic resonance (MR) imaging). The use of these techniques has dramatically increased in recent years, despite severe criticism of their validity and effectiveness, and practice guidelines advising against the routine use of such techniques [6,9–11]. Studies evaluating the validity and reliability of different lumbosacral spine image reading methods have reported mixed results, ranging from low to high levels of validity and reliability [12–16]. Taken together, all of this suggests that diagnosing low back pain is a highly complex task. Establishing new means of improving diagnostic accuracy for image reading in the context of low back pain is thus imperative.

One hitherto unexplored mechanism for increasing diagnostic accuracy of spine image reading is to apply a collective intelligence approach. Collective intelligence refers to the phenomenon that multiple minds can solve cognitive tasks better than single minds [17–22]. Collective intelligence can arise via different mechanisms, such as group discussions with consensus seeking (e.g., Delphi-technique, nominal group technique) but also by algorithmically combining independent decisions. Here we will focus on the latter approach. Previous work has shown the potential of such an approach in increasing the diagnostic accuracy of dermatologists evaluating skin lesions [23,24], radiologists evaluating mammograms [24,25], clinicians predicting positive bone scans [26], and medical students diagnosing simulated patients arriving at the emergency room [27]. However, the extent to which collective intelligence could improve diagnostic accuracy in the case of difficult-to-diagnose low back pain is currently unknown.

In this study, we investigated the benefits of utilizing a collective intelligence approach as a means of increasing diagnostic accuracy of interpreting lumbosacral radiographs and MR images. The objective was to compare the performance of the majority rule (a powerful collective intelligence rule which combines the independent decisions of multiple raters) against the performance of single raters, both in terms of sensitivity and specificity. We used two different individual benchmarks: the performance of the average single individual in a group, and the performance of the best single individual in a group.

Materials and methods

We used two data-sets from two previously-published studies ([15,16] see also Fig 1 and S1 File). We briefly describe both studies below and refer for further methodological details to the original studies due to space constraints.

Study 1 radiographs

Study 1 investigated the reliability of lumbosacral spine radiograph reading by chiropractors (n = 5), chiropractic radiologists (n = 3) and medical radiologists (n = 5) (for rater details see [15]). Each of the 13 raters independently rated a set of 300 patient-blinded lumbosacral radiographs. Radiographs were derived from a database from the Medical Centrum Alkmaar (the Netherlands). Only radiographs that involved the entire lumbar vertebrae and more than half of the sacrum and from patients greater than or equal to 18 years of age were included (for detailed description of image selection see [15]). The study was designed to investigate the reliability of detecting "significant abnormalities", defined as conditions which have a major influence on the continued well-being of a patient. When detected, these conditions warrant immediate referral to a hospital or the intervention needs to be modified. The significant abnormalities included: infection (n = 7), malignancy (n = 15), fracture (n = 8), inflammatory spondylitis (n = 6) and spondylolysis-spondylolisthesis (n = 14). 50 out of 300 radiographs used in the study thus had a significant abnormality (i.e., prevalence: 16.7%). The reference

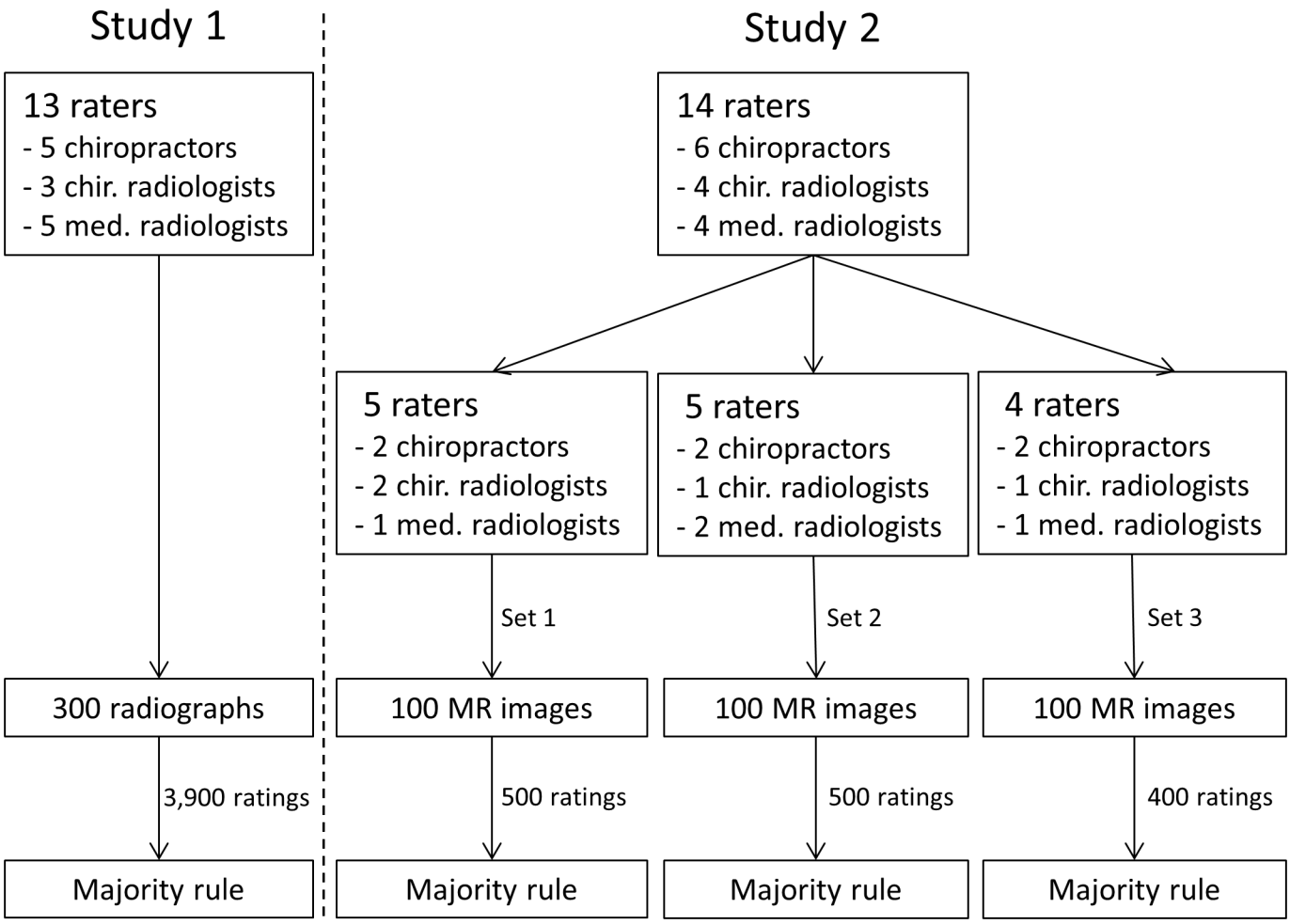

Fig 1. Flow diagram showing the procedures of both studies. In study 1, all 13 raters evaluated all 300 radiographs. In study 2, five raters evaluated 100 MR images belonging to set 1; five raters evaluated 100 MR images belonging to set 2; and four raters evaluated 100 MR images belonging to set 3. Chir. radiologists = Chiropractic radiologists. Med. radiologists = Medical radiologists.

https://doi.org/10.1371/journal.pone.0194128.g001

test was based on clinical findings, if present lab data, and/or MR/CT imaging. A chiropractor and medical expert in spinal radiology checked that the abnormality was detectable and of sufficient quality.

All of the 13 raters were fully licensed from their professional organizations. Each rater independently evaluated the set of 300 radiographs twice, three months apart. For the purpose of this study, we consider only the first session of evaluations. Before the rating started, a meeting was organized to ensure uniform interpretation of the rating list. For each radiograph, a rater indicated whether the radiograph in question contained an "abnormality" or not. Raters were not informed about the total number of abnormalities present. For further experimental details of the study see [15].

Study 2 MR images

Study 2 investigated the reliability of lumbosacral spine MR image reading by chiropractors (n = 6), chiropractic radiologists (n = 4) and medical radiologists (n = 4) (for full rater details see [16]). Each rater was assigned to one of three image sets, each containing 100 unique,

patient-blinded MR images (sets 1 & 2 were evaluated by five raters and set 3 by four raters, respectively).

Three hundred MR images of the lumbosacral spine of patients referred by primary care clinicians and specialists were selected retrospectively from the Medical Centrum Alkmaar (the Netherlands). Only images from patients greater than or equal to 18 years of age and of sufficient image quality were selected (for detailed description of image selection see [16]). The reference test was based on the evaluations by one experienced chiropractic radiologist and one experienced medical radiologist (who were not raters in the actual study). If they disagreed, a third medical radiologist had the final say. All three experts were specialized in musculoskeletal imaging and they had substantially more years of experience (respectively, 8, 19 and 26) than the 14 raters (median number of years' experience of raters: 6 years). Nonetheless, the use of an expert panel as reference test is not optimal. To investigate this issue further, we studied the number of cases in which the expert panel disagreed on whether the image should be labelled as specific finding or not (see below). The experts disagreed on six images [16]. Therefore, we also analysed our results while excluding these six ambiguous cases. We further revisit this issue in the discussion.

Specific findings were defined as infections, malignancies, fractures, herniated disc and central stenosis. In case of such specific findings, patients need to be referred to a medical specialist or the treatment needs to be modified. Two classifications were used, because the criteria to define nerve root involvement in disc herniation and central stenosis are not always unambiguous and because we used an expert panel as reference test. In classification 'A' infection, malignancy, fracture, herniated disc with definitive root involvement and central stenosis with definite nerve root involvement were classified as "specific finding". In classification 'B' herniated disc and central stenosis with doubtful nerve root involvement were also classified as "specific finding". Each of the three image sets had approximately equal prevalence of specific findings (prevalence classification 'A': set 1: 32%, set 2: 30%, set 3: 31%; classification 'B': set 1: 57%, set 2: 56%, set 3: 57%). The image sequence was randomized differently for each rater.

Each rater rated each set of MR images on the presence of the five specific findings. These were then dichotomized in "abnormal" and "normal". As in Study 1, raters evaluated MR images in two sessions. For the purpose of this study, we consider only the first session of evaluations. For further details of the study see [16].

For both studies, the medical ethical committee of the Alkmaar hospital approved the study. Potential participants were approached via an email, containing information about the study. This was followed up by a phone call to the potential participants in which the study was explained in more detail. Participants who then expressed verbal consent (that is, after receiving full information about the study) were enrolled in the study.

In sum, Study 1 comprised 300 lumbosacral spine radiographs all rated by 13 raters (3,900 ratings in total). Study 2 comprised three sets of 100 unique lumbosacral spine MR images, whereby set 1 and 2 were evaluated by five raters, and set 3 by four raters (1,400 ratings in total) (see also Fig 1). This allowed us to test how combining independent ratings affected collective accuracy using a majority rule.

Majority rule versus the average individual performance

We investigated the performance of the majority rule, which chooses as final diagnosis the diagnosis receiving most support. The majority rule is a powerful method for boosting collective accuracy under a wide range of individual accuracy levels and decision making contexts [28–31]. To test how group size affected collective accuracy, we tested a range of group sizes. For study 1, we tested group sizes ranging from 3 to 9; for study 2, we tested group sizes

ranging from 3 to 5. We could not test larger group sizes in Study 2 because raters only evaluated the MR images within their own test set (Fig 1). Furthermore, we only evaluated odd group sizes to avoid the need of a tie-breaker rule. Within each study, and for study 2 within each test set, we randomly drew groups of *n* raters. We then pooled all the ratings of the *n* raters and evaluated for each radiograph (study 1) / MR image (study 2) whether more raters were in favor of 'normal' (in which case the diagnostic image was labeled normal) or 'abnormal' (in which case the diagnostic image was labelled as abnormal). From this, we calculated the performance of the majority rule in terms of (i) sensitivity (i.e., the percentage of abnormal cases correctly diagnosed as abnormal), (ii) specificity (i.e., the percentage of normal cases correctly diagnosed as normal), and (iii) the Youden's index (*J*). The Youden's index combines sensitivity and specificity in one diagnostic measure [32,33] and is calculated as: J = sensitivity + specificity– 1. A perfect test has $J = 1$ (i.e., sensitivity = specificity = 1), whereas a test with no discriminatory power has $J = 0$ (i.e., sensitivity = 1—specificity; that is, the test has equal probability of giving a positive result for both normal and abnormal cases). Next, we calculated for each group the average individual performance using the same three performance measures. We then compared the performance of the majority rule to the average individual performance of the same group (i.e., performance majority rule minus average individual performance). For study 1, we repeated this procedure 250 times for each group size (i.e., 3, 5, 7 and 9). We compared whether the resulting distributions were statistically higher than zero (i.e., the null hypothesis that there is no difference between the performance of the majority rule and the average individual group performance) at the $P < 0.05$ level. For 250 groups, this implies that the number of groups at or below zero should not exceed $\lfloor 250 \cdot 0.05 = \rfloor$ 12. For study 2, we could not perform the same number of simulations because each of the raters evaluated only one out of three image sets. Therefore, we created the maximum number of unique groups possible at each group size, which was 24 for group size 3, and only two for group size 5. To be statistically significant at the $P < 0.05$ level at group size 3, the number of groups with values at or below zero should not exceed $\lfloor 24 \cdot 0.05 = \rfloor$ 1. For group size 5, the number of unique groups was limited to two. Therefore, we were not able to draw any statistical inference for this group size. Additionally, we used a permutation test for paired samples (comparing the performance of the majority rule with the average individual performance of that group) to verify our findings. This was done using the coin package in R (version 3.2.2).

Finally, we tested whether group size affected improvement in the three performance measures (i.e., Youden's index, sensitivity and specificity) using the full range of group sizes tested. We used general linear models (LMs), using group size (continuous) as a fixed effect and the improvement in the performance measure as response variable, running a different model for each performance measure. This analysis was done in R (version 3.4.0) and significance levels were derived from the t-values and associated p-values.

Majority rule versus the best individual

Next, we investigated under which conditions combining the decisions of raters allowed the collective to outperform the best individual rater in a group. Thus, in this case we did not take the average individual performance of a group as the benchmark but instead the performance of the best rater in a given group. We specifically investigated how differences in average individual performance among group members affected the ability of a group to outperform the best single rater in that group. This was done because recent work in visual perception tasks found that combining decisions of individuals of similar individual performance level resulted in a collective performance which was better than any single individual [34–37]. However, when individuals substantially differ in their average individual performance level, combining

their decisions leads to worse performance level as compared to the performance of the best individual. (Similar results were recently found in breast and skin cancer diagnostics [24].) Since both the similarity between raters in a group and the identity of the best rater in a group need to be known before they can be used (in practice), we performed a cross validation procedure. First, within each study, we created all possible combinations of raters at group size 3. For each group, we randomly assigned 70% of all images to a training set (210 and 70 images for study 1 and 2 respectively) and the remaining 30% to a test set (90 and 30 images for study 1 and 2). The training set was used to (i) determine the identity of the best performing rater of the triplet in terms of the Youden's index, and (ii) determine the similarity in accuracy between group members. For this, we calculated the Youden's index for each group member in the training set and used the mean pairwise absolute deviation (MPAD) to calculate the similarity in J among group members.

$$\mathrm{MPAD} = \frac{2}{n(n-1)} * \sum_{i<j} |J_i - J_j|,$$

where n is the number of raters i and j. This measure is thus the expected absolute difference in J between two randomly chosen group members. Next, we determined the performance of the best individual (selected from the training set) in the test set, as well as the performance of the majority rule in the test set. Performance was again measured in terms of (i) sensitivity, (ii) specificity, and (iii) the Youden's index. We repeated this procedure 500 times for each unique group composition, and averaged results within unique groups. We then studied how the "similarity in accuracy" (estimated from a training set) affected the performance of the majority rule in the test set as compared to the performance of the best rater (also selected from the training set) in the test set. For this, we analyzed the effect of "similarity in accuracy" on a group's ability to outperform its best individual using LMs in R. Significance levels were derived from the t-values and associated p-values.

Results

Majority rule versus the average individual performance

Fig 2 shows the performance of the majority rule compared to the average individual performance of that group for study 1 (Radiographs). At each group size, 250 simulations were performed. For all group sizes, the majority performance in terms of the Youden's index of all the 250 simulated groups was better than the average individual performance of that group (i.e., none of the distributions overlapped with zero) (Fig 2A). This provides statistical evidence in favor of a significant difference from zero at all group sizes. For sensitivity, the 95% CIs at group sizes 3 and 5 overlapped with zero and were thus not significantly different from zero (Fig 2B). However, for group sizes 7 and 9, the distributions were significantly greater than zero (Fig 2B). For specificity, none of the distributions overlapped with zero (Fig 2C). The permutation tests for paired samples largely confirmed these findings: all $P < 0.01$ for each combination of group size (3, 5, 7 and 9) and performance measure (Youden's index, sensitivity and specificity).

Fig 3 shows the absolute performance of the different group sizes, illustrating how all three performance measures (Youden's index, sensitivity and specificity) increased with increasing group size (results of LMs using full range of group sizes: all $P < 0.01$). The largest improvements in all measures arose when the number of raters increased from one to three, illustrating that the largest gains were obtained at the lower end of the group size range.

Study 2 consisted of three test sets (each containing 100 unique MR images), rated by respectively five, five and four raters. Within each test set, we evaluated the performance of the

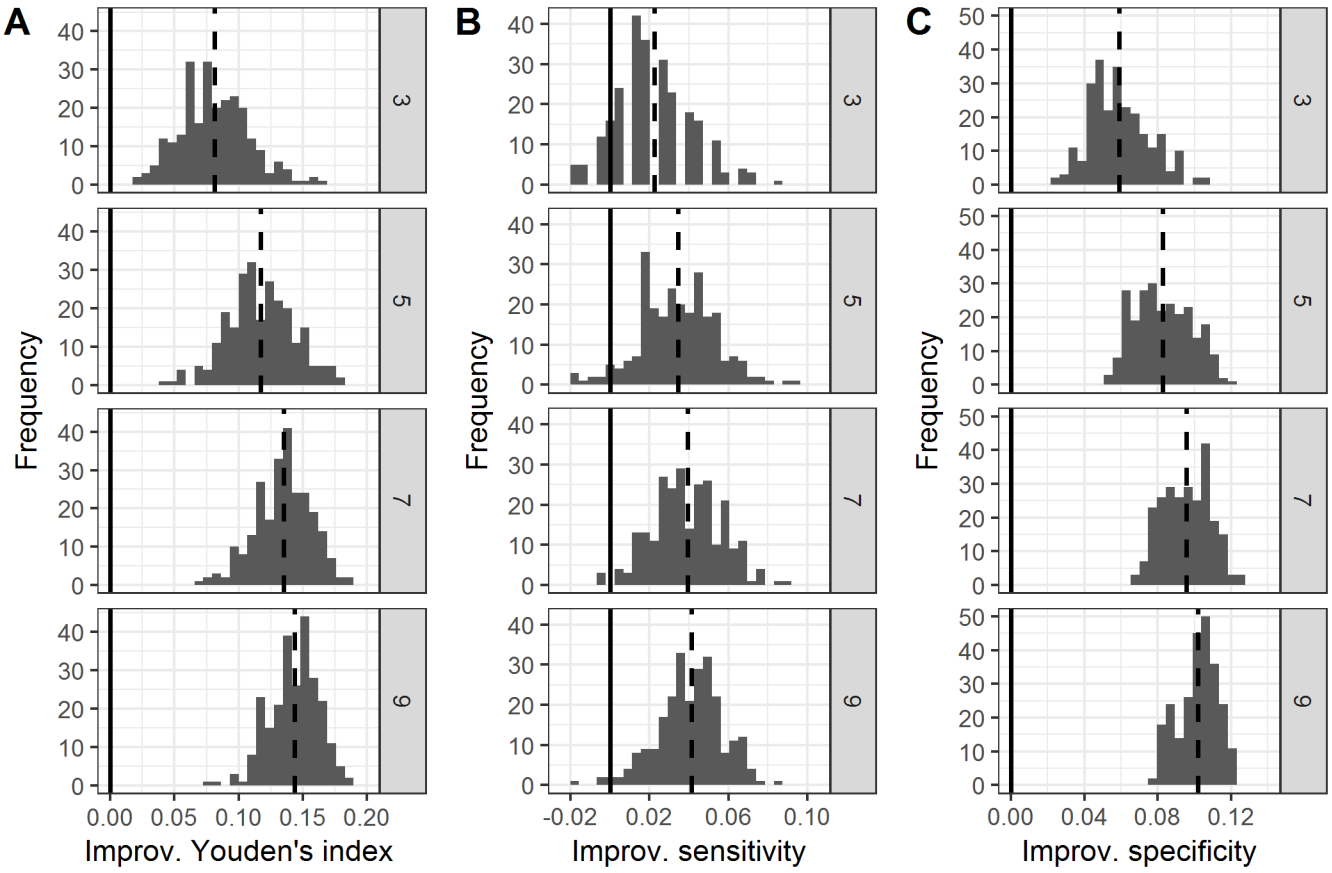

Fig 2. Effect of group size on the Youden's index, sensitivity and specificity for reading lumbosacral spine radiographs (study 1). Histograms show the frequency distributions of the improvement of groups under the majority rule as compared to the average individual performance of that group, in terms of (A) the Youden's index, (B) sensitivity, and (C) specificity. At each group size (numbers in grey panels), 250 unique groups were drawn. Values higher than zero indicate that the majority rule was better than the average individual performance of that group. Negative values indicate that the majority rule was worse than the average individual performance of that group. The dashed vertical lines show the mean value of each distribution. The solid vertical lines represent the average individual group performance (which by definition corresponds to an improvement of zero). Improv = Improvement.

https://doi.org/10.1371/journal.pone.0194128.g002

majority rule using classification A. Fig 4 shows the performance of the majority rule compared to the average individual performance of that group, for group sizes 3 (24 unique groups) and 5 (only 2 unique groups). At group size 3, the majority performance in terms of the Youden's index of all the 24 simulated groups was better than the average individual performance of that group (i.e., the distribution did not overlap with zero) (Fig 4A). Furthermore, the improvements in sensitivity (Fig 4B) and specificity (Fig 4C) at group size 3 were also significantly greater than zero. The permutation tests for paired samples confirmed these findings: $P < 0.01$ for all three performance measures (Youden's index, sensitivity and specificity).

Fig 5 shows the absolute performance of the different group sizes, for each of the three image sets. For all sets, applying the majority rule consistently increased the Youden's index (Fig 5A–5C), sensitivity (Fig 5D–5F) and specificity (Fig 5G–5I) (results of LMs: all $P < 0.01$). Despite substantial differences in average individual sensitivity/specificity between the three test sets, we found that in all tests increasing group size consistently increased both sensitivity and specificity. We found similar results when looking at ratings using classification B (S1 Fig). Finally, we also found similar results (for classifications A and B) when excluding the six cases for which the expert panel could not find consensus (S2 Fig).

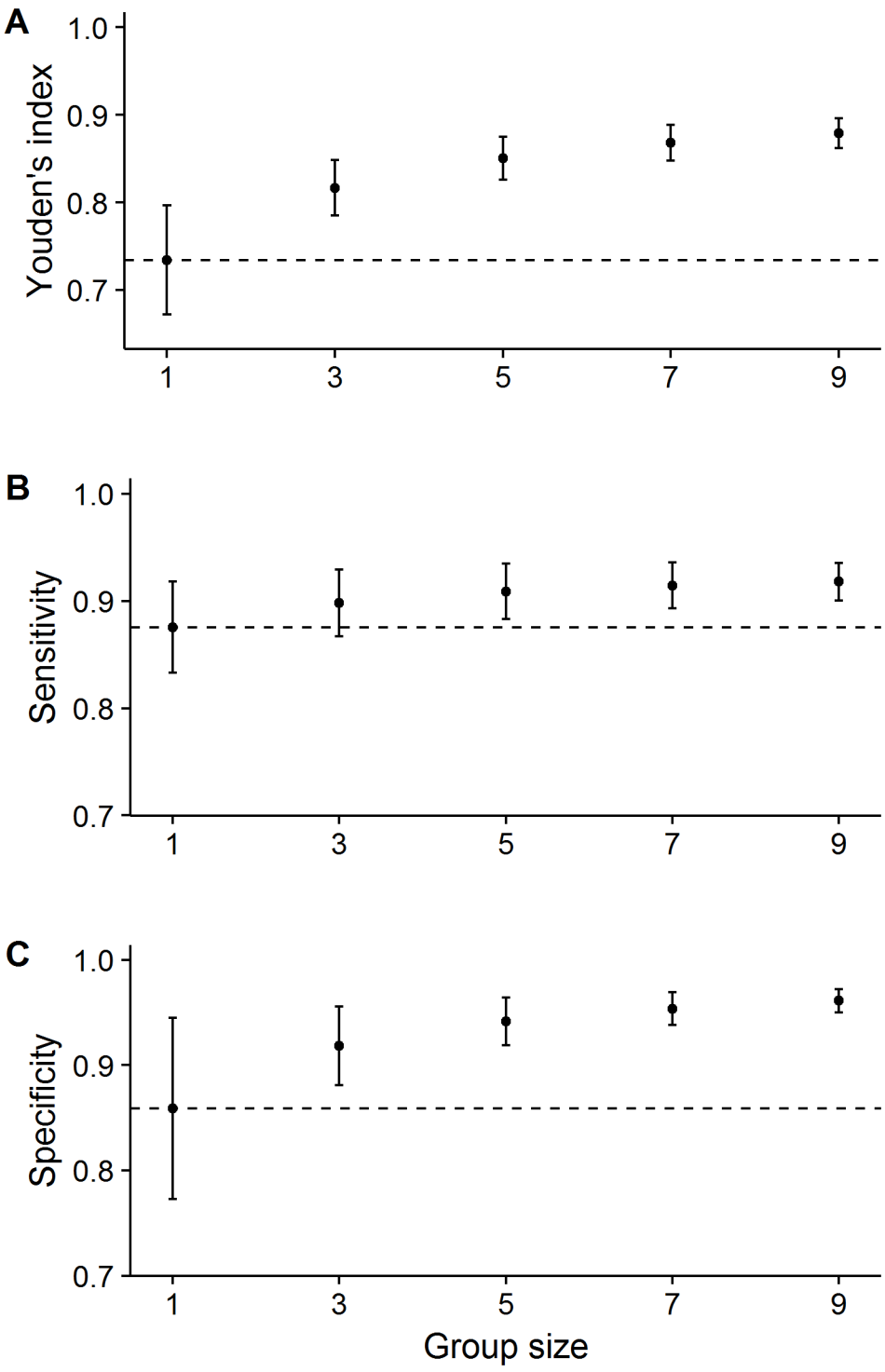

Fig 3. Effect of group size on the absolute performance in terms of the Youden's index, sensitivity and specificity for reading lumbosacral spine radiographs (study 1). Increasing the number of independent ratings under the majority rule increased (A) the Youden's index, (B) sensitivity and (C) specificity. Horizontal lines show the average individual performance (i.e., group size = 1). Error bars represent standard deviation.

https://doi.org/10.1371/journal.pone.0194128.g003

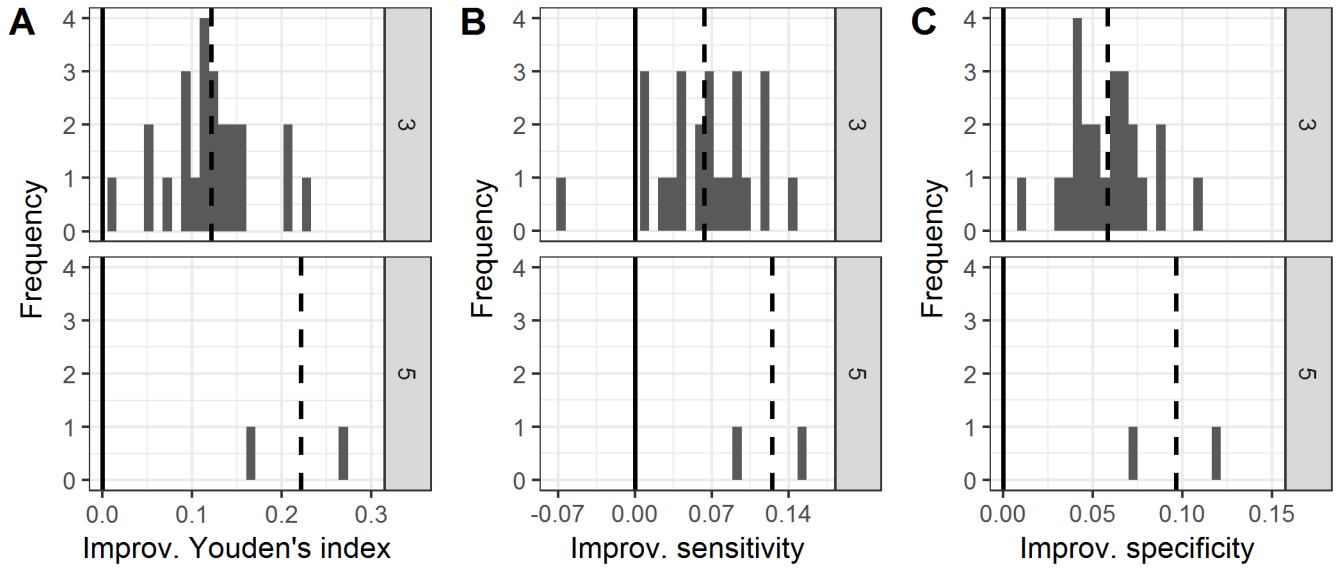

Fig 4. Effect of group size on the Youden's index, sensitivity and specificity for reading lumbosacral spine MR images (study 2). Histograms show the frequency distributions of the improvement of groups under the majority rule as compared to the average individual performance of that group, in terms of (A) the Youden's index, (B) sensitivity, and (C) specificity. At group size three, 24 unique groups were available, and at group size five, two unique groups. Values higher than zero indicate that the majority rule was better than the average individual performance of that group. Negative values indicate that the majority rule was worse than the average individual performance of that group. The dashed vertical lines show the mean value of each distribution. The solid vertical lines represent the average individual group performance (which by definition corresponds to an improvement of zero). Improv = Improvement.

https://doi.org/10.1371/journal.pone.0194128.g004

Majority rule versus the best individual

In both studies, we found that as the difference in performance levels among the three raters (in the training set) increased, the performance of the majority rule (in the set) as compared to the best rater (in the test set) decreased (Youden's index: results of LM, Study 1: estimate (est) ± se = -0.80 ± 0.06, t = -12.81, $p < 0.001$, Fig 6A; Study 2: est ± se = -1.63 ± 0.27, t = -5.97, $p < 0.001$, Fig 6B). When raters had relatively similar individual performance level (i.e., $\Delta J < 0.17$), combining their decisions under the majority rule led to better decisions as compared to the best single rater (Fig 6A and 6B). In contrast, when raters were relatively dissimilar in individual performance level ($\Delta J > 0.17$), combining their decisions led to worse decisions as compared to the best single rater. When separating the overall performance (i.e., Youden's index) into sensitivity and specificity (Fig 6C–6F), we observe that the negative relationship between performance difference and the ability of a group to outperform the best rater, is driven by specificity in Study 1 (Fig 6E) but by sensitivity in Study 2 (Fig 6D). (LM, Study 1: sensitivity: est ± se = 0.05 ± 0.05, t = 0.98, $p > 0.3$; specificity: est ± se = -0.85 ± 0.04, t = -22.3, $p < 0.001$; Study 2: sensitivity: est ± se = -1.39 ± 0.43, t = -3.20, p = 0.004; specificity: est ± se = -0.23 ± 0.27, t = -0.88, $p > 0.3$).

Discussion

This study showed that pooling independent ratings increases the diagnostic accuracy in lumbosacral spine image interpretation: increasing the number of independent ratings increased both sensitivity and specificity. These results were found in both studies, each one using a different imaging technique (i.e., radiographs and MR images respectively). Our results corroborate earlier findings in different domains of medical diagnostics which have shown an increased diagnostic accuracy when pooling independent diagnostic decisions in radiology

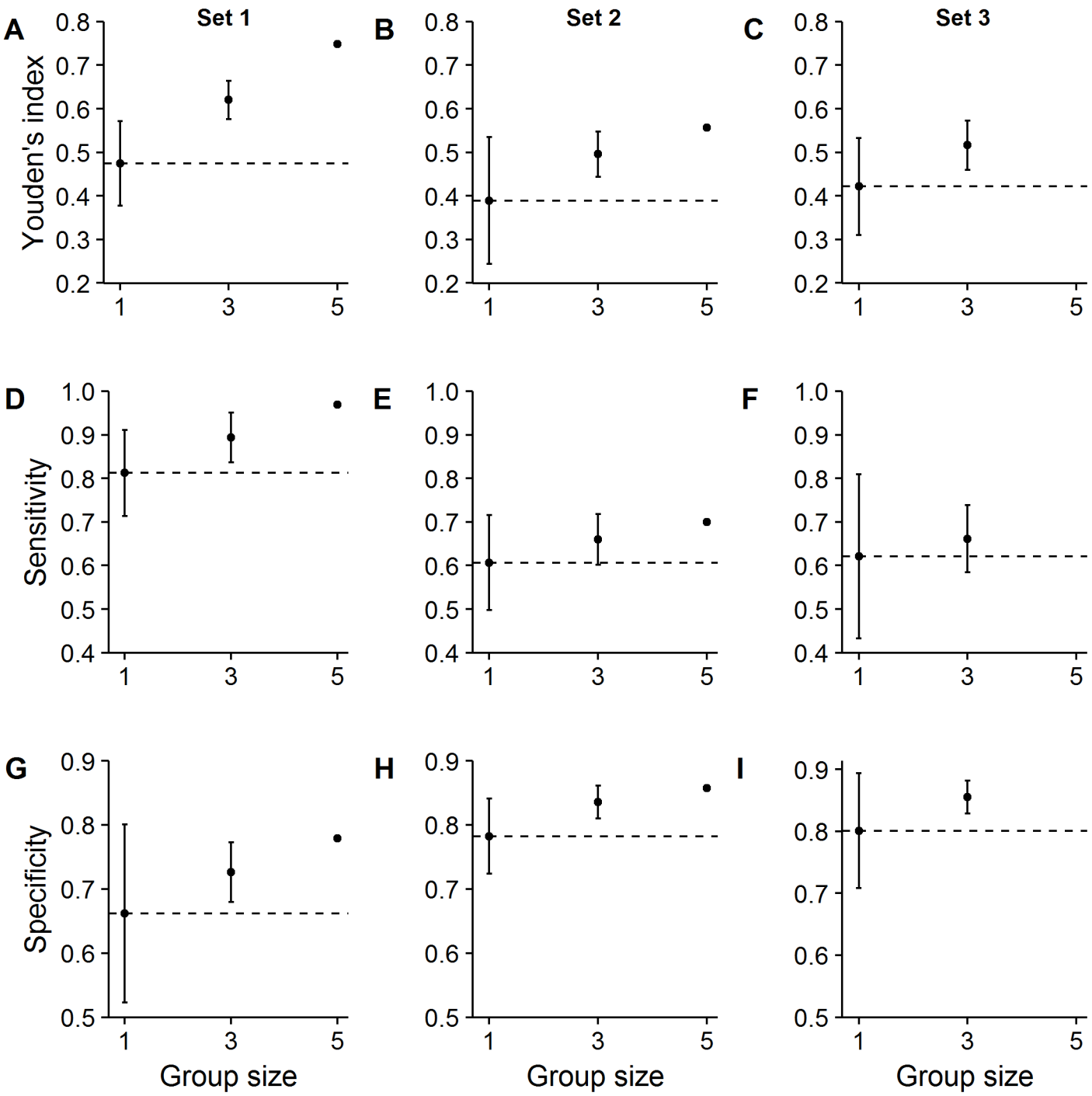

Fig 5. Effect of group size on the absolute performance in terms of the Youden's index, sensitivity and specificity for reading lumbosacral spine MR images (study 2). Increasing the number of independent ratings under the majority rule increased (A-C) the Youden's index, (D-F) sensitivity and (G-I) specificity in all three test sets. Horizontal lines show the average individual performance (i.e., group size = 1). Error bars represent standard deviation (S.D.). Each test set contained 100 unique MR images rated by respectively five, five and four raters, explaining why group size five in test set 3 is absent, and the absence of S.D. for group size five in test set 1 and 2 (since there was only one unique combination of five raters).

https://doi.org/10.1371/journal.pone.0194128.g005

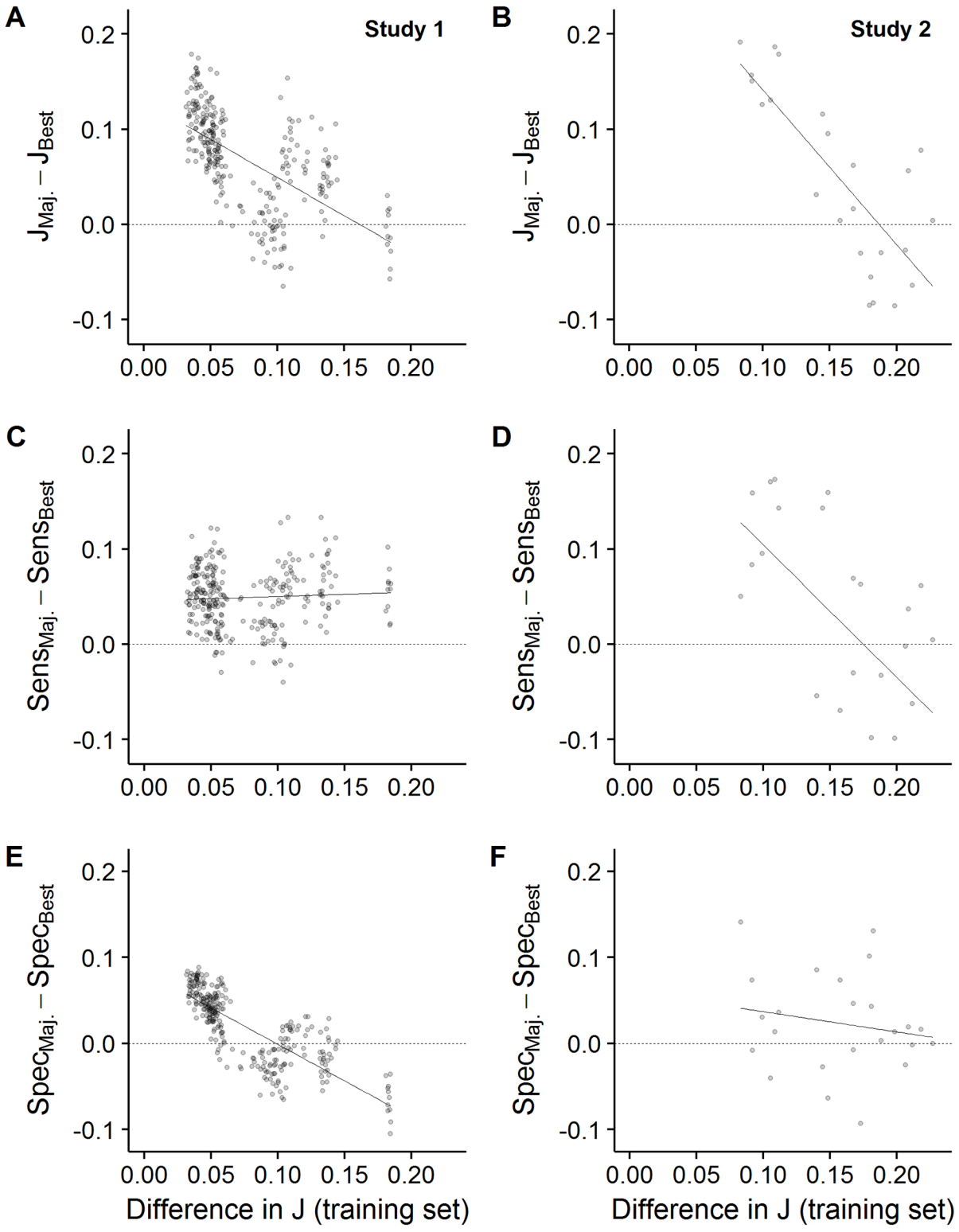

Fig 6. Performance difference between the majority rule and the best rater in a group of three raters as a function of the difference in performance level (i.e., difference in Youden's index) between raters. Each point represents a unique combination of three raters. Values above zero indicate that combining independent ratings under the majority rule outperformed the best rater in that group by that many percentage points in terms of (A, B) the Youden's index (*J*) (C, D) sensitivity (sens) and (E, F) specificity (spec). Values below zero indicate that the best rater outperformed the majority rule. Lines are linear regression lines. Data are shown separately for (A, C, E) Study 1 and (B, D, F) Study 2. Results show groups averages based on a cross validation procedure with 500 repetitions per unique group composition (see Methods and materials for more details).

https://doi.org/10.1371/journal.pone.0194128.g006

[24,25], dermatology [23,38] positive bone scan predictions [26] and emergency medicine [27].

Although groups in general outperformed the average performance of single raters, whether groups also outperformed the best rater in a group depended critically on the similarity in performance level among raters (Fig 6A and 6B). When raters were of similar performance, combining their ratings resulted in outcomes which were better than those of any of the group members. However, when group members differed too much in their individual performance level, combining their decisions resulted in worse outcomes as compared to the best rater. In practice, this implies that if there is no prior information on average individual performance levels, then combining ratings/decisions is to be preferred (as combining outperforms the average performance). However, whenever prior information on rater's performance is available, then this could be used to determine whether combining should be preferred over the best individual, or to combine specific raters of similar individual performance. Our separate treatments of sensitivity and specificity (Fig 6C–6F) show that considering these two key dimensions of performance can be important when doing this, since these performance measures might scale differently with rater similarity.

Another method for combining the decisions of multiple raters is to conduct a group discussion followed by a joint group decision. Group discussions have been shown to increase performance in several domains [39–42], but at the same time several studies have highlighted the pitfalls associated with group discussions, including social loafing, group think and obedience to authority [43–45]. We currently do not know how our mechanism of combining independent decisions compares to scenarios with group discussions followed by a joint group decision, and future research could pitch these collective mechanisms against each other to compare the potential collective gains of both methods [46]. Future studies investigating combining independent ratings in low back diagnoses could also collect data on the confidence raters have in their decision which would allow testing other, more complex, collective intelligence rules such as (weighted) confidence rules [24,34,35,47] which give more weight to highly confident decisions. Future studies would also benefit from a higher number of independent raters. In our simulations, we created unique groups, but these groups consisted (partly) of the same raters, introducing dependence between groups. In an ideal scenario, more raters would be available to avoid such dependencies. Simultaneously, obtaining a large sample of medical experts can be a challenge so such ideal scenarios have to be traded off against practical feasibility.

The costs and benefits of using lumbosacral spine imaging for patients with low back pain is heavily debated and current guidelines advise against the use of routine imaging. Only in the presence of progressive neurological deficits or symptoms suggesting a serious or specific underlying condition, is the use of imaging recommended [3,7]. Our study did not directly address the reliability and validity of imaging under different patient specific scenarios. Our study does, however, show that when imaging is used, there is scope for improving the diagnostic accuracy by combining independent decisions of raters. Importantly, this improvement arose in the two key dimensions of diagnostic accuracy: sensitivity and specificity. Improvements in one dimension thus did not go at the expense of improvements on the other dimension, as for example was the case in some studies investigating the diagnostic accuracy of second opinions in mammography. Several of these studies reported that second opinions increased sensitivity (as compared to single decision-makers) but at the expense of decreased specificity [48–50].

One important limitation of our study constitutes the reference tests used. Although Study 1 used a combination of clinical findings, if present lab data, MR/CT imaging, and an expert panel, the reference test in study 2 was solely based on a consensus expert panel. An important

assumption we made is thus that the ratings of the expert panel correlate with the truth. The gold standard for identifying serious underlying pathologies in low back pain remains a challenge, as there is a lack of a gold reference standard. In the literature, there is a large variation of reference tests. To illustrate, in a Cochrane review on the diagnostic accuracy of MRI on low back pain, studies were only included if they used surgery, expert panel consensus or diagnostic work up as reference standard. Surgery, especially when combined with clinical follow-up, is often regarded as the best reference test, but subject to partial verification as often only patients with a strong suspicion of a specific underlying cause will be subjected to surgery. Verification bias might lead to a higher sensitivity and a lower specificity but it has also been found that it increases both sensitivity and specificity [51]. More studies are needed to identify best possible imaging strategies in patients with chronic low-back pain, symptoms of radiculopathy or spinal stenosis, patients assessed in referral settings, and other specific subgroups [6]. Another limitation of our study is that in both data-sets we analyzed, the prevalence of "abnormalities" was substantially higher than in clinical practice, implying that the collective gains we found cannot be directly transferred to clinical populations. Future studies using more realistic prevalence rates could investigate the consequences for clinical populations directly. Despite the elevated prevalence rates, it is noteworthy that in all the data-sets we investigated, combining decisions improved sensitivity and specificity; despite there being substantial differences in average individual sensitivity and specificity levels (e.g., Fig 5). This suggests that our findings are, at least partly, independent of the exact average sensitivity and specificity level of raters. A further important consideration of our study is that applying a collective intelligence approach requires more viewing time by health care providers. These additional costs need to be weighed against the potential gains: increased sensitivity (i.e., higher detection rates of serious abnormalities) and increased specificity (potentially lowering costly unnecessary or even harmful follow-up treatments). Future studies could quantify how to optimize benefits of a collective intelligence approach while balancing costs.

Conclusions

Our findings suggest that employing a collective intelligence approach can improve both sensitivity and specificity of lumbosacral diagnostic imaging reading. These results were found in two different studies using different imaging methods (i.e., radiographs and MR images).

Supporting information

S1 Fig. Effect of group size on the Youden's index, sensitivity and specificity for reading lumbosacral spine MR images (study 2) using classification B. Histograms show the frequency distributions of the improvement of groups under the majority rule as compared to the average individual performance of that group, in terms of (A) the Youden's index, (B) sensitivity, and (C) specificity. At group size three, 24 unique groups were available, and at group size five, two unique groups. Values higher than zero indicate that the majority rule was better than the average individual performance of that group. Negative values indicate that the majority rule was worse than the average individual performance of that group. The dashed vertical lines show the mean value of each distribution. The solid vertical lines represent the average individual group performance (which by definition corresponds to an improvement of zero). Improv = Improvement. At group size three, the majority performance was significantly better than the average individual performance in terms of the Youden's index and sensitivity, but not in terms of specificity.
(TIFF)

S2 Fig. Effect of group size on the Youden's index, sensitivity and specificity for reading lumbosacral spine MR images (study 2) using classification A and B when excluding the six ambiguous cases for which the expert panel could not find consensus. Histograms show the frequency distributions of the improvement of groups under the majority rule as compared to the average individual performance of that group, in terms of (A, D) the Youden's index, (B, E) sensitivity, and (C, F) specificity. At group size three, 24 unique groups were available, and at group size five, two unique groups. Values higher than zero indicate that the majority rule was better than the average individual performance of that group. Negative values indicate that the majority rule was worse than the average individual performance of that group. The dashed vertical lines show the mean value of each distribution. The solid vertical lines represent the average individual group performance (which by definition corresponds to an improvement of zero). Improv = Improvement. At group size three, the majority performance was significantly better than the average individual performance in all six panels, except for specificity under classification B.
(TIFF)

S1 File. Data sets Kurvers et al. 2018. The two data sets used in this study.
(XLSX)

Acknowledgments

The authors thank two anonymous reviewers for their valuable comments on earlier versions of this manuscript. The authors thank C. Peterson, M. Wessely and J.T. Wilmink for their help with the selection of the radiographs and MRI; C.E. Janssen, S.F.C. Knaap, J.A. Kuipers, G.A. Regelink, K.L. Schott, P. Byrne, I. Dijkers, T. Soeters, L. van Viegen, T. van der Hof, W. de Koning, (chiropractors), A. Thorkeldsen, A.H. Turner, J.A. Wylie-Cook, C. Tao, A. Manne, J. Cooley, M. Fergus (chiropractic radiologists), R. Hagenbeek, H. van Woerden, C. Hoeberigs, W.P.M. Klazen, K.J. Simon, W.R. Oberman, G.M.J.M. Vanderschueren and H. J. Teertstra (medical radiologists) for the many hours they spent as assessors.

Author Contributions

Conceptualization: Ralf H. J. M. Kurvers, Shelby L. Bachman.

Data curation: Ralf H. J. M. Kurvers, Annemarie de Zoete, Shelby L. Bachman, Paul R. Algra, Raymond Ostelo.

Formal analysis: Ralf H. J. M. Kurvers, Shelby L. Bachman.

Funding acquisition: Raymond Ostelo.

Investigation: Ralf H. J. M. Kurvers, Annemarie de Zoete, Shelby L. Bachman, Paul R. Algra, Raymond Ostelo.

Resources: Annemarie de Zoete, Paul R. Algra, Raymond Ostelo.

Visualization: Ralf H. J. M. Kurvers, Shelby L. Bachman.

Writing – original draft: Ralf H. J. M. Kurvers.

Writing – review & editing: Annemarie de Zoete, Shelby L. Bachman, Paul R. Algra, Raymond Ostelo.

References

1. Deyo RA, Mirza SK, Martin BI (2006) Back pain prevalence and visit rates: estimates from US national surveys, 2002. Spine 31: 2724–2727. https://doi.org/10.1097/01.brs.0000244618.06877.cd PMID: 17077742
2. Hart LG, Deyo RA, Cherkin DC (1995) Physician office visits for low back pain: frequency, clinical evaluation, and treatment patterns from a US national survey. Spine 20: 11–19. PMID: 7709270
3. Chou R, Deyo R, Jarvik J (2012) Appropriate Use of Lumbar Imaging for Evaluation of Low Back Pain. Radiologic Clinics of North America 50: 569-+. https://doi.org/10.1016/j.rcl.2012.04.005 PMID: 22643385
4. Luo X, Pietrobon R, Sun SX, Liu GG, Hey L (2004) Estimates and patterns of direct health care expenditures among individuals with back pain in the United States. Spine 29: 79–86. https://doi.org/10.1097/01.BRS.0000105527.13866.0F PMID: 14699281
5. Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingworth W, et al. (2008) Expenditures and health status among adults with back and neck problems. JAMA 299: 656–664. https://doi.org/10.1001/jama.299.6.656 PMID: 18270354
6. Chou R, Fu R, Carrino JA, Deyo RA (2009) Imaging strategies for low-back pain: systematic review and meta-analysis. The Lancet 373: 463–472.
7. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Shekelle P, et al. (2007) Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. Ann Intern Med 147: 478–491. PMID: 17909209
8. Cherkin DC, Deyo RA, Wheeler K, Ciol MA (1994) Physician variation in diagnostic testing for low back pain. Who you see is what you get. Arthritis Rheum 37: 15–22. PMID: 8129759
9. Baras JD, Baker LC (2009) Magnetic resonance imaging and low back pain care for Medicare patients. Health Aff (Millwood) 28: w1133–w1140.
10. Deyo RA, Mirza SK, Turner JA, Martin BI (2009) Overtreating chronic back pain: time to back off? The Journal of the American Board of Family Medicine 22: 62–68. https://doi.org/10.3122/jabfm.2009.01.080102 PMID: 19124635
11. Tan A, Zhou J, Kuo Y-F, Goodwin JS (2016) Variation among primary care physicians in the use of imaging for older patients with acute low back pain. J Gen Intern Med 31: 156–163. https://doi.org/10.1007/s11606-015-3475-3 PMID: 26215847
12. Carrino JA, Lurie JD, Tosteson AN, Tosteson TD, Carragee EJ, Kaiser J, et al. (2009) Lumbar Spine: Reliability of MR Imaging Findings 1. Radiology 250: 161–170. https://doi.org/10.1148/radiol.2493071999 PMID: 18955509
13. Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino J, Kaiser J, et al. (2008) Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. Spine 33: 1605. https://doi.org/10.1097/BRS.0b013e3181791af3 PMID: 18552677
14. Deyo RA, Mirza SK (2016) Herniated Lumbar Intervertebral Disk. N Engl J Med 374: 1763–1772. https://doi.org/10.1056/NEJMcp1512658 PMID: 27144851
15. de Zoete A, Assendelft WJ, Algra PR, Oberman WR, Vanderschueren GM, Bezemer PD (2002) Reliability and validity of lumbosacral spine radiograph reading by chiropractors, chiropractic radiologists, and medical radiologists. Spine 27: 1926–1933. PMID: 12221360
16. de Zoete A, Ostelo R, Knol DL, Algra PR, Wilmink JT, van Tulder MW, et al. (2015) Diagnostic Accuracy of Lumbosacral Spine Magnetic Resonance Image Reading by Chiropractors, Chiropractic Radiologists, and Medical Radiologists. Spine 40: E653–E660. https://doi.org/10.1097/BRS.0000000000000896 PMID: 25803219
17. Krause J, Ruxton GD, Krause S (2010) Swarm intelligence in animals and humans. Trends Ecol Evol 25: 28–34. https://doi.org/10.1016/j.tree.2009.06.016 PMID: 19735961
18. 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: 20122777.
19. 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: 686–688. https://doi.org/10.1126/science.1193147 PMID: 20929725
20. 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.
21. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford: Oxford University Press.

22. Kurvers RH, Wolf M, Naguib M, Krause J (2015) Self-organized flexible leadership promotes collective intelligence in human groups. Royal Society open science 2: 150222. https://doi.org/10.1098/rsos.150222 PMID: 27019718

23. Kurvers RHJM, Krause J, Argenziano G, Zalaudek I, Wolf M (2015) Detection accuracy of collective intelligence assessments for skin cancer diagnosis. JAMA Dermatol 151: 1–8.

24. Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, et al. (2016) Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A 113: 8777–8782. https://doi.org/10.1073/pnas.1601827113 PMID: 27432950

25. 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: e0134269. https://doi.org/10.1371/journal.pone.0134269 PMID: 26267331

26. Kattan MW, O'Rourke C, Yu C, Chagin K (2016) The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones. Med Decis Making 36: 536–540. https://doi.org/10.1177/0272989X15581615 PMID: 25878196

27. Kammer JE, Hautz WE, Herzog SM, Kunina-Habenicht O, Kurvers RHJM (2017) The Potential of Collective Intelligence in Emergency Medicine: Pooling Medical Students' Independent Decisions Improves Diagnostic Performance. Med Decis Making 37: 715–724. https://doi.org/10.1177/0272989X17696998 PMID: 28355975

28. Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. Psychol Rev 112: 494–508. https://doi.org/10.1037/0033-295X.112.2.494 PMID: 15783295

29. Grofman B, Owen G, Feld SL (1983) Thirteen theorems in search of the truth. Theor Decis 15: 261–278.

30. Boland PJ (1989) Majority systems and the Condorcet Jury Theorem. Statistician 38: 181–189.

31. Sorkin RD, West R, Robinson DE (1998) Group performance depends on the majority rule. Psychol Sci 9: 456–463.

32. Hilden J, Glasziou P (1996) Regret graphs, diagnostic uncertainty and Youden's index. Stat Med 15: 969–986. https://doi.org/10.1002/(SICI)1097-0258(19960530)15:10<969::AID-SIM211>3.0.CO;2-9 PMID: 8783436

33. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3: 32–35. PMID: 15405679

34. Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD (2010) Optimally interacting minds. Science 329: 1081–1085. https://doi.org/10.1126/science.1185718 PMID: 20798320

35. Koriat A (2012) When are two heads better than one and why? Science 336: 360–362. https://doi.org/10.1126/science.1216549 PMID: 22517862

36. Bang D, Fusaroli R, Tylen K, Olsen K, Latham PE, Lau JYF, et al. (2014) Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. Conscious Cogn 26: 13–23. https://doi.org/10.1016/j.concog.2014.02.002 PMID: 24650632

37. Koriat A (2015) When Two Heads Are Better Than One and When They Can Be Worse: The Amplification Hypothesis. J Exp Psychol Gen 144: 934–950. https://doi.org/10.1037/xge0000092 PMID: 26168039

38. King AJ, Gehl RW, Grossman D, Jensen JD (2013) Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches. Cancer Epidemiol 37: 979–984. https://doi.org/10.1016/j.canep.2013.09.004 PMID: 24075797

39. Kerr NL, Tindale RS (2004) Group performance and decision making. Annu Rev Psychol 55: 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009 PMID: 14744229

40. Clément RJG, Krause S, von Engelhardt N, Faria JJ, Krause J, Kurvers RHJM (2013) Collective cognition in humans: Groups outperform their best members in a sentence reconstruction task. PLoS ONE 8: e77943. https://doi.org/10.1371/journal.pone.0077943 PMID: 24147101

41. Klein N, Epley N (2015) Group discussion improves lie detection. P Natl Acad Sci USA 112: 7460–7465.

42. Hackman JR, Morris CG (1975) Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. Advances in experimental social psychology 8: 45–99.

43. Blass T (1999) The Milgram Paradigm after 35 years: Some things we now know about obedience to authority. J Appl Soc Psychol 29: 955–978.

44. Turner ME, Pratkanis AR (1998) Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. Organ Behav Hum Decis Process 73: 105–115. PMID: 9705798

45. McGrath JE (1984) Groups: Interaction and performance: Prentice-Hall: Englewood Cliffs, NJ.

46. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. Psychol Assess 12: 19. PMID: 10752360

47. Moussaïd M, Yahosseini KS (2016) Can simple transmission chains foster collective intelligence in binary-choice tasks? PLoS ONE 11: e0167223. https://doi.org/10.1371/journal.pone.0167223 PMID: 27880825

48. Gromet M (2008) Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. Am J Roentgenol 190: 854–859.

49. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J (2001) Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. The Breast 10: 455–463. https://doi.org/10.1054/brst.2001.0350 PMID: 14965624

50. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B (2003) Increase in cancer detection and recall rates with independent double interpretation of screening mammography. Am J Roentgenol 180: 1461–1467.

51. Wassenaar M, van Rijn RM, van Tulder MW, Verhagen AP, van der Windt DA, Koes BW, et al. (2012) Magnetic resonance imaging for diagnosing lumbar spinal pathology in adult patients with low back pain or sciatica: a diagnostic systematic review. Eur Spine J 21: 220–227. https://doi.org/10.1007/s00586-011-2019-8 PMID: 21922287