The Potential of Collective Intelligence in Emergency Medicine: Pooling Medical Students' Independent Decisions Improves Diagnostic Performance

Juliane E. Kämmer, PhD, Wolf E. Hautz, MD, MME, Stefan M. Herzog, PhD, Olga Kunina-Habenicht, PhD, Ralf H. J. M. Kurvers, PhD

Background. *Evidence suggests that pooling multiple independent diagnoses can improve diagnostic accuracy in well-defined tasks. We investigated whether this is also the case for diagnostics in emergency medicine, an ill-defined task environment where diagnostic errors are rife.* ***Methods.*** *A computer simulation study was conducted based on empirical data from 2 published experimental studies. In the computer experiments, 285 medical students independently diagnosed 6 simulated patients arriving at the emergency room with dyspnea. Participants' diagnoses (n = 1,710), confidence ratings, and expertise levels were entered into a computer simulation. Virtual groups of different sizes were randomly created, and 3 collective intelligence rules (follow-the-plurality rule, follow-the-most-confident rule, and follow-the-most-senior rule) were applied to combine the independent decisions into a final diagnosis. For different group sizes, the performance levels (i.e., percentage of correct diagnoses) of the 3 collective intelligence rules were compared with each other and against the average individual accuracy.* ***Results.*** *For all collective intelligence rules, combining independent decisions substantially increased performance relative to average individual performance. For groups of 4 or fewer, the follow-the-most-confident rule outperformed the other rules; for larger groups, the follow-the-plurality rule performed best. For example, combining 5 independent decisions using the follow-the-plurality rule increased diagnostic accuracy by 22 percentage points. These results were robust across case difficulty and expertise level. Limitations of the study include the use of simulated patients diagnosed by medical students. Whether results generalize to clinical practice is currently unknown.* ***Conclusion.*** *Combining independent decisions may substantially improve the quality of diagnoses in emergency medicine and may thus enhance patient safety.* ***Key words:*** *collective intelligence; wisdom of crowds; medical diagnostics; emergency medicine; simulation; follow-the-plurality rule.* ***(Med Decis Making 2017;37:715–724)***

Every year, approximately 250,000 people in the United States alone die from preventable medical errors,[1] many of them diagnostic errors.[2–5] In addition, incorrect diagnoses substantially contribute to incorrect treatment and patient morbidity,[6–8] especially in emergency medicine.[9] Despite its crucial importance, research on effective strategies to decrease diagnostic errors is still limited.[6,10–12] Here, we test whether combining independent diagnoses can improve diagnostic accuracy in virtual decision scenarios in emergency medicine. Emergency medicine is a complex decision environment where many initial diagnoses are made and diagnostic errors are rife.[13]

For many judgments, pooling independent decisions can outperform the average and sometimes even the best individual,[14] because different individuals' errors can cancel out at the group level.[15,16] This "wisdom-of-crowds"[14,17] phenomenon has gained importance in a variety of domains, such as business, economics, and politics, and has long been studied in social psychology under the term "statisticized groups."[18–22] Previous studies have demonstrated the potential of such a collective intelligence approach also in clinical diagnostics, but these studies have examined well-defined tasks such as interpreting mammograms or diagnosing

DOI: 10.1177/0272989X17696998

skin lesions,[23–28] where the diagnosis is based on 1 or a few pieces of readily available information, there are few time constraints, and the decision is binary (e.g., cancer/no cancer). In emergency medicine, by contrast, diagnoses are often based on incomplete information and made under severe time pressure, and diagnosticians must consider many competing diagnoses.[29,30] Crucially, these diagnostic decisions can have immediate and severe consequences for patients.

Patients arriving at the emergency room (ER) with acute conditions are frequently seen by 2 or more physicians. Whereas noncritical patients are typically seen by at least 1 junior and 1 senior physician, team size can increase up to 15 clinicians or more[31–34] in high-acuity situations.[3] Physicians then make a joint diagnosis in an informal, nonstandardized way. Such team decisions may be based, for example, on independent examinations by each physician involved or on collaborative examinations (if a team assembles at the bedside). In addition, asking for a second (or further) opinion(s) may be either unconditional or conditional on the first physician's assessment. In case of disagreement, physicians may initiate discussion and reach an informal consensus or the decision may be referred to, for example, the most senior physician. Alternatively, several independently made diagnoses could be pooled in a standardized, algorithmic way. Here, we evaluate the performance of this latter approach by studying the performance of different collective intelligence rules that combine several independently, unconditionally made diagnoses into 1 final diagnosis using virtual decision scenarios from emergency medicine.

Received 22 June 2016 from the Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany (JEK, SMH, RHJMK); AG Progress Test Medizin, Charité Medical School, Berlin, Germany (JEK); Department of Emergency Medicine, Inselspital, University Hospital Bern, Bern, Switzerland (WEH); and German Institute for International Educational Research, Centre for International Student Assessment, Frankfurt am Main, Germany (OKH). This work was presented previously at the 58th Conference of Experimental Psychologists, 21–23 March 2016, Heidelberg, Germany; and the 50th Conference of the German Society for Psychology, 18–22 September 2016, Leipzig, Germany. Revision accepted for publication 30 January 2017.

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm.

Address correspondence to Juliane E. Kämmer, PhD, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; e-mail: kaemmer@mpib-berlin.mpg.de.

METHODS

Our analyses are based on previously published data sets from 2 studies[35,36] that are described below.

Experimental Procedures

Both studies assessed clinical reasoning using the norm-referenced computer-based Assessing Clinical Reasoning (ASCLIRE) test.[36] The experimental task was to diagnose simulated patients with dyspnea, a common symptom in the ER.[37] Six patients with acute or subacute dyspnea, each with a different correct diagnosis (see Supplemental Table S1), were presented in random order.

The interpretation of ASCLIRE test scores has been internally and externally validated[36] against established frameworks of validity.[38,39] Test cases based on real patients were developed by 3 board-certified anesthesiologists, 2 board-certified internists, and 2 educational psychologists. Case selection was comparable to that of other studies assessing clinical reasoning,[40] suggesting good agreement on representativeness and relevance for the domain of dyspnea. In detail, we used an expert-based consensus method to identify relevant cases, whereby relevant was defined as frequent, urgent, or exemplary. We then identified real patients diagnosed with the respective conditions and incorporated their data into the test (i.e., their X-rays, their electrocardiograms [ECGs], and so on). A patient actor was trained to enact the symptoms of these patients and audio responses were recorded from this simulated patient, not the real patient.

The simulated patient presented prototypical symptoms for the given disease, as indicated by the near-perfect performance of an expert panel of 8 anesthesiologists and 12 internists with considerable context-relevant professional expertise (mean expert accuracy 94.2%).[36] A study assessing 283 medical students (from the first to fifth year of study) with this test revealed a monotone increase in diagnostic accuracy across year of study, with significant differences between years with new relevant curricular content,[36] which is considered an important criterion to assess a test's validity.[41] The same study showed that experts had higher accuracy rates and needed less time than students, both typical findings distinguishing experts from novices.[42–44] Furthermore, diagnostic performance in this test, which was designed to assess aspects of clinical reasoning, did not correlate with other performance measures such as factual or

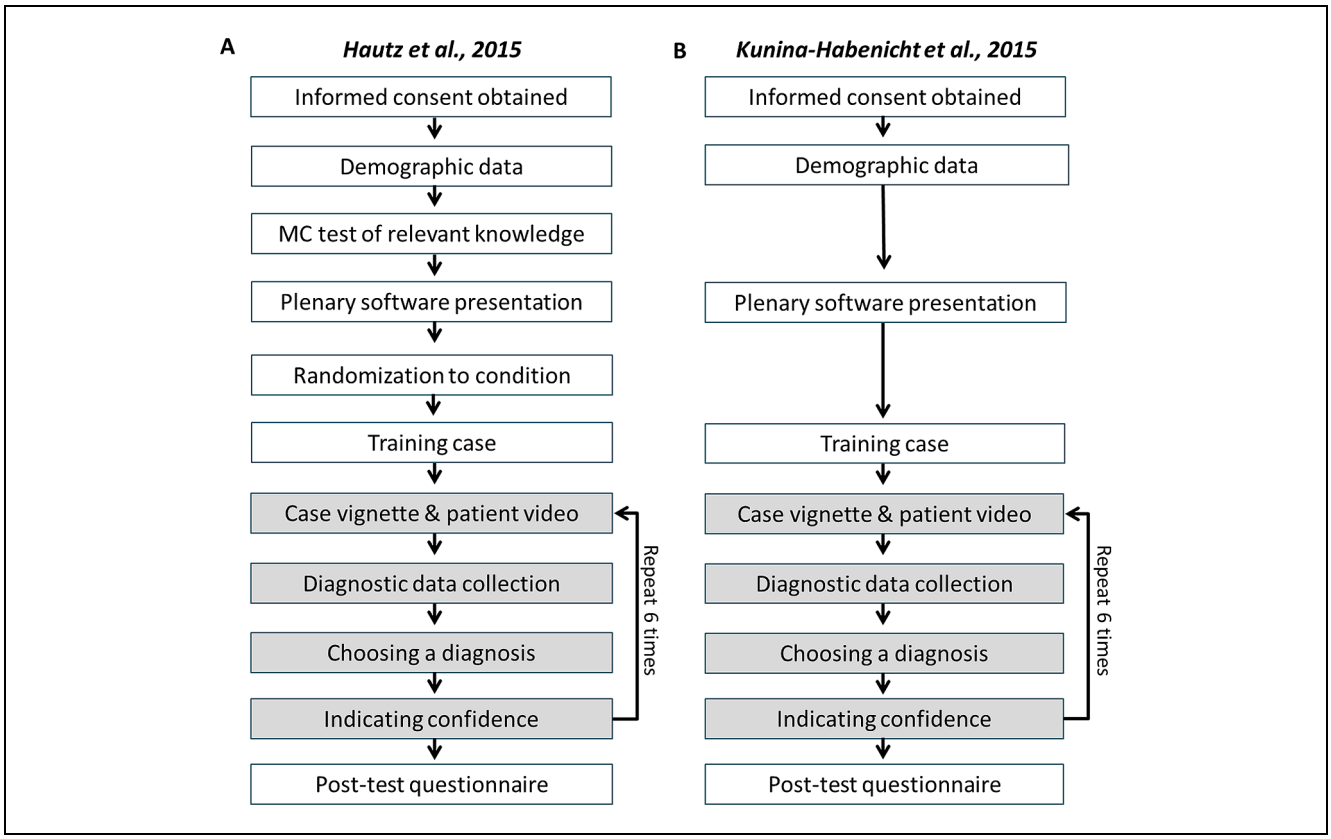

Figure 1 (A and B) General procedures used in the 2 experimental studies by Hautz et al.[35] (A) and Kunina-Habenicht et al. (B).[36] Note that study A also contained a second condition with interacting pairs (not included in the current reanalysis) and thus a phase in which participants were randomly allocated to conditions. MC, multiple choice.

procedural knowledge.[36] Latent reliability Ω ranged between 0.63 and 0.85, depending on the measure of performance.[36]

Figure 1 summarizes the procedure of the ASCLIRE test in the 2 experimental studies.[35,36] Participants first read the study descriptions and signed a consent form. After answering demographic questions, participants received a demonstration of how to work on the clinical test cases. After a training case, they worked individually on the 6 randomly ordered test cases. (To note, one study[36] additionally used a second condition with interacting pairs; these data are, however, not used here because, by design, those diagnoses were not produced independently.) Per case, participants first watched a short video clip showing the same male, standardized actor patient with case-specific prototypical symptoms and makeup. For each case, participants then collected patient-specific information using a graphical interface on the computer screen (Figure 2; also see Table 1 in Kunina-Habenicht et al.[36] for details). Participants were free to choose any type, order, and number of diagnostic tests, the results of which were displayed via text (e.g., pulse rate), image (e.g., ECG, chest X-ray), or audio (e.g., heart sounds, history) and had to be interpreted by the participant. This need to acquire and interpret the diagnostic test results renders the ASCLIRE a more high-fidelity test than the frequently used multiple-choice examinations.

Participants were instructed to work as fast as possible without sacrificing accuracy. Participants eventually had to choose 1 of 20 possible diagnoses (or 1 of 3 "other" options; see Supplemental Table S1); the set of possible diagnoses was the same across all cases and was known to participants from the training case. After choosing their lead diagnosis, participants indicated their level of confidence in their diagnosis on a 10-point Likert scale ranging from least confident to most confident. The testing procedure allowed, in principle, for several differential diagnoses to be evaluated by students before deciding on the most likely one and reporting their confidence in this diagnosis.

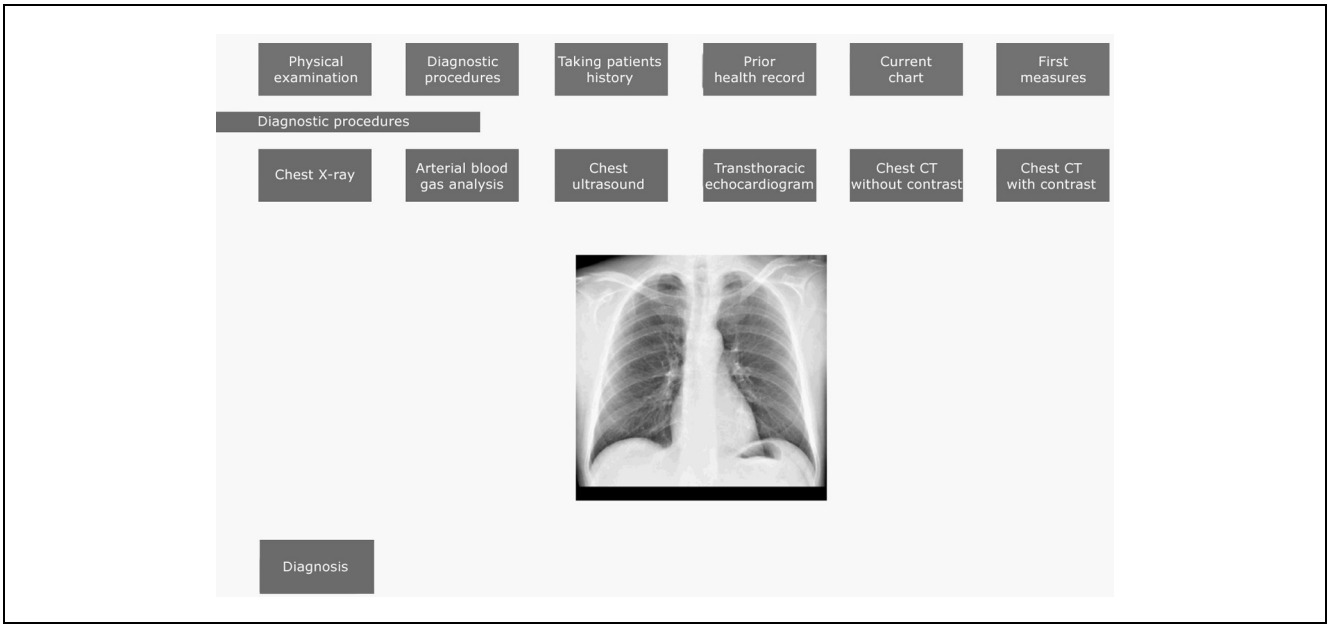

Figure 2 Screenshot of the experimental task. When participants selected 1 of the 6 main categories (first row), subcategories appeared in the second row, which could again be selected, resulting in the diagnostic test result of that particular subcategory (here, an X-ray of the chest). Participants were free to provide a diagnosis (left bottom corner) at any time. The screenshot is translated into English from German. CT, computed tomography scan.

Table 1 Demographics and Diagnostic Accuracy of Participants in the Final Sample

Origin	Year	No. of Participants in the Subsample	Age, Years	Gender, % Women	Proportion Correct
Hautz et al., 2015[35]	1	—	—	—	—
	2	2	23.0 (0.0)	50.0	66.7 (0.0)
	3	5	27.2 (2.8)	80.0	46.7 (27.4)
	4	17	24.0 (1.8)	70.6	43.1 (25.5)
	5	3	24.7 (1.2)	69.6	72.2 (25.5)
	Total	28[a]	24.8 (2.4)	67.9	50.0 (24.4)
Kunina-Habenicht et al., 2015[36]	1	56	22.4 (5.5)	69.6	36.6 (19.7)
	2	55	23.9 (6.0)	72.7	48.2 (24.6)
	3	44	24.6 (5.4)	70.5	55.7 (21.2)
	4	48	25.9 (4.4)	62.5	69.4 (20.4)
	5	54	26.4 (4.4)	63.0	71.0 (18.7)
	Total	257	24.6 (5.4)	67.7	55.7 (24.7)
Total sample	1	56	22.4 (4.6)	69.6	36.6 (19.7)
	2	57	24.0 (5.9)	71.9	48.8 (24.4)
	3	49	24.9 (5.2)	71.4	54.8 (21.8)
	4	65	25.4 (4.0)	64.6	62.6 (23.8)
	5	57	26.3 (4.3)	63.2	71.1 (18.8)
	Total	285[a]	24.6 (5.1)	67.0	55.1 (24.7)

Note: Values are given as means (SDs) unless otherwise indicated.
[a]One participant did not indicate her study year.

This is very close to clinical practice where patients are admitted to the hospital or discharged from the ER with typically one lead diagnosis. If a student used a differential-diagnoses approach and could not exclude all other diagnoses, this would (or should) presumably lead to low-confidence diagnoses.

In both studies, the complete session lasted about 2 h. In 1 study,[35] participants were compensated with €25 ($33 at that time); in the other study,[36] participation was part of the curriculum and not compensated.

Participants

Participants were 311 medical students (68% female, 32% male, study years 1–5) from the Charité Medical School in Berlin, Germany (283 participants from the Kunina-Habenicht et al.[36] study and 28 participants from the Hautz et al.[35] study). All participants had clinical experience, either because they were already advanced in their studies or because they were enrolled in a reformed medical curriculum (in the study by Kunina-Habenicht et al.[36]), which provides heavy clinical exposure from the first day onward.[45]

For 26 of those participants, 1 or more confidence responses were not recorded due to a technical error. These participants were excluded from all analyses, resulting in a final sample of 285 participants (see Table 1 for demographics). The 285 × 6 = 1,710 individually rendered diagnoses were entered into the computer simulation, without any student-specific variables except for their study year and confidence ratings. The Charité Medical School institutional review board approved both studies (under EA 1/170/09 and EA 1/276/12).

Accuracy Criteria

We used 2 complementary accuracy criteria (Supplemental Table S1): diagnostic accuracy and treatment adequacy. First, diagnostic accuracy indicated whether or not the diagnosis was correct. Second, since not all diagnostic errors in the ER are equally severe, we also evaluated treatment adequacy. Seven experts independently categorized (for each case) all 19 incorrect diagnoses as either ''implied treatment partially adequate'' or ''implied treatment not adequate'' (i.e., whether or not the treatment implied by the diagnosis would have been at least adequate and not harmful, given the true diagnosis; see below for an example). Experts were board-certified consultants in emergency medicine with at least 10 years of professional experience, all currently working as supervising physicians in a level I ER. Experts' interrater agreement across cases was moderately high (mean Fleiss kappa ± SE = 0.57 ± 0.02; range, 0.51–0.61). We used a majority rule to aggregate expert ratings of a diagnosis as implying adequate or inadequate treatment. Treatment adequacy is arguably more subjective than diagnostic accuracy; therefore, we focus mainly on diagnostic accuracy. However, treating all incorrect decisions as equally severe (as diagnostic accuracy does) neglects a crucial aspect of decisions in emergency medicine—namely, that some incorrect decisions are worse than others. For example, misdiagnosing a pulmonary embolism as a myocardial infarction would—although the diagnosis is wrong—still imply thrombolytic therapy together with the application of oxygen and monitoring and/or pharmaceutically supporting cardiac output, whereas misdiagnosing a pulmonary embolism as pneumonia would lead to antibiotic treatment and likely to fluid restriction, both unnecessary or even harmful for patients with a pulmonary embolism.

Collective Intelligence Rules

To test the performance of a collective intelligence approach, we randomly created groups of different sizes (n = 2–15, 20, or 25), applied 3 different collective intelligence rules to all 6 cases, and compared the groups' diagnostic accuracy against that of the average individual (Figure 3). We deliberately included very large group sizes of up to 25 as benchmarks to study how well small groups already approximate the collective intelligence of much larger groups. Groups were created using computer simulations, and group members thus did not

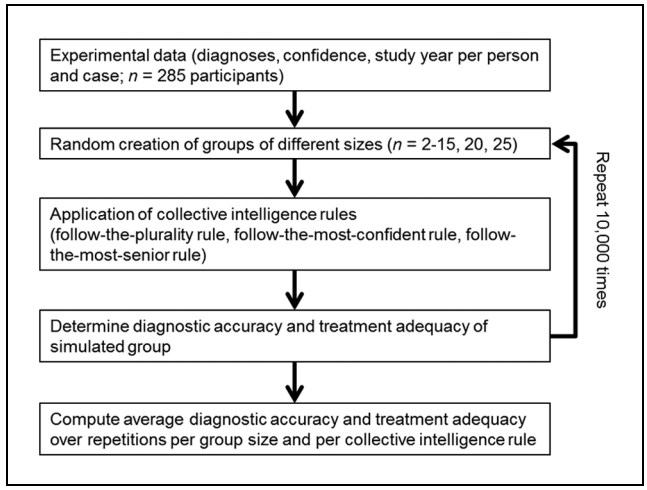

Figure 3 General procedure of the collective intelligence simulations.

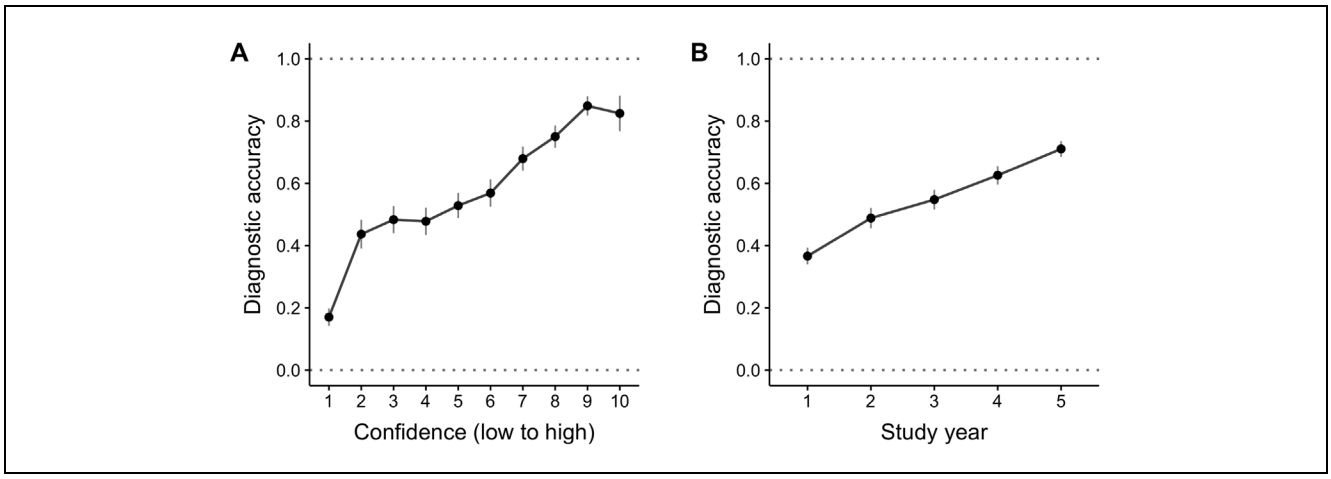

Figure 4 (A and B) Mean diagnostic accuracy per confidence rating (A) and study year (B). Error lines represent standard errors of the mean proportion correct across participants.

interact. Per group size, we ran 10,000 repetitions (in MATLAB R2014b; MathWorks, Natick, MA, USA) using the following collective intelligence rules:

1. The follow-the-plurality rule picks the diagnosis chosen by most group members[46]; in case of a tie, 1 of the tied diagnoses was randomly selected. The follow-the-plurality rule was applied from a group size of 3 upward.
2. The follow-the-most-confident rule picks the diagnosis with the highest confidence rating[47–49]; in case of a tie, 1 of the tied diagnoses was randomly selected. This rule performs well when confidence and accuracy correlate positively, which was the case (Figure 4A).
3. The follow-the-most-senior rule picks the diagnosis of the most senior group member (in terms of study year); if group members with the same highest seniority selected different diagnoses, 1 of those diagnoses was randomly selected. We used study year as a proxy for level of expertise.[36,50,51] This rule performs well when seniority and accuracy correlate positively, which was the case (Figure 4B).

In summary, the results presented here involve high-fidelity simulated patient profiles, which were independently examined by medical students; 3 collective intelligence rules were applied to computer-simulated, virtual teams of different sizes. The study had no external funding source.

RESULTS

For all collective intelligence rules, combining independent decisions increased diagnostic accuracy (Figure 5A) and substantially outperformed the average individual. For example, considering a second opinion and following the decision with the higher confidence rating increased diagnostic accuracy by 10 percentage points. Combining 5 independent decisions using the follow-the-plurality rule increased diagnostic accuracy by 22 percentage points. Overall, the follow-the-most-confident rule slightly outperformed the follow-the-most-senior rule. Beyond a group size of 4, both rules were outperformed by the follow-the-plurality rule. Moreover, the follow-the-plurality rule continued to improve as group size increased up to 25, whereas the gains achieved by the other 2 rules quickly leveled off.

Although the 6 cases differed greatly in difficulty (range of average individual diagnostic accuracy, 36–78%; Supplemental Figure S1), the collective intelligence rules improved accuracy in all cases, suggesting that this approach is appropriate for both easy and difficult cases (Supplemental Figure S2A). Similarly, although individuals' diagnostic accuracy increased with seniority (from 37% for first-year to 71% for fifth-year medical students; Figure 4B and Table 1), the collective intelligence rules improved accuracy within each seniority level (Supplemental Figure S3A).

We found parallel effects for treatment adequacy: Increasing group size increased treatment adequacy

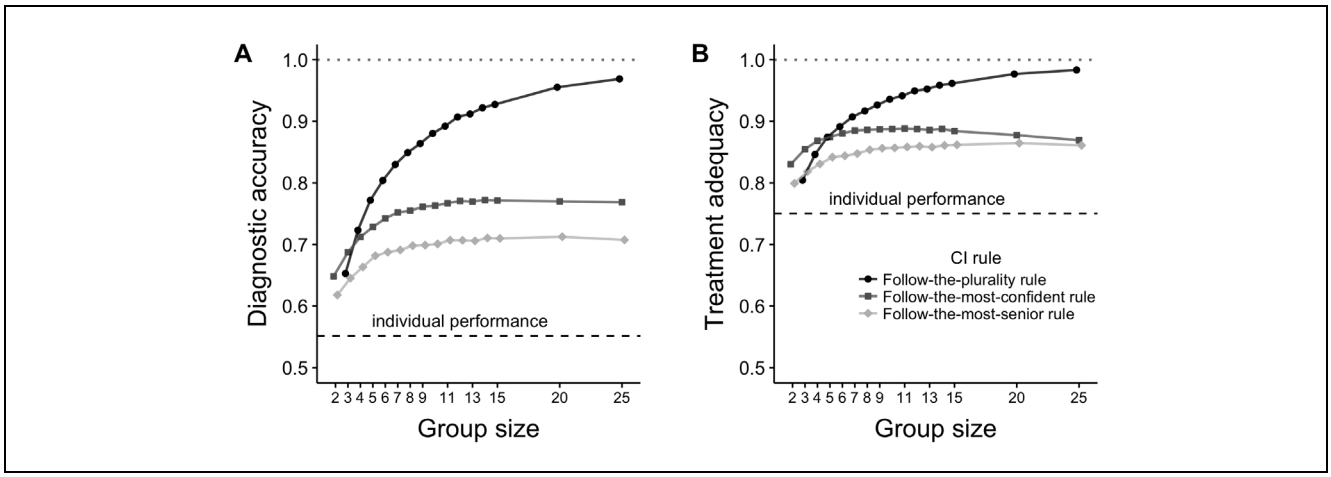

Figure 5 Performance of the 3 collective intelligence rules averaged across all 6 cases. (A and B) For all 3 rules, increasing the number of independent decisions increased diagnostic accuracy (A) and treatment adequacy (B). CI, collective intelligence.

(Figure 5B) for all collective intelligence rules, with the follow-the-plurality rule outperforming both other rules beyond a group size of 5. Similar results were obtained for the 6 cases separately (Supplemental Figure S2B), and results for the follow-the-plurality and follow-the-most-confident rules were roughly invariant across all seniority levels (Supplemental Figure S3B).

DISCUSSION

Misdiagnosis is one of the greatest concerns for patients in outpatient and hospital settings,[52] has important legal and economic consequences,[30] and can severely affect patients' health,[1,12] especially in emergency medicine.[9] Our findings suggest that combining independent decisions could substantially improve diagnostic accuracy for patients arriving at the ER with dyspnea, a common and difficult-to-diagnose symptom.[37] Importantly, our results were invariant across case difficulty, expertise level (study year), and 2 complementary accuracy criteria; that is, the collective benefits were also observed for difficult cases and senior students. Although these results are based on high-fidelity simulated patients examined by medical students (see below for a more detailed discussion of the limitations of our study), they nevertheless suggest that a collective intelligence approach has the potential to increase diagnostic accuracy and thus also patient safety and to reduce inadequate treatments in the domain of emergency medicine and other ill-defined task environments where decisions have to be based on incomplete information and a large choice set has to be considered. Our results encourage the explicit use of collective intelligence rules in settings where patients are regularly seen by more than 1 diagnostician. In the ER, for example, patients are often seen by a surprisingly large number of clinicians ranging up to 15 or even more.[31–34] Combining independent diagnoses in an algorithmic way could be a time-saving alternative to traditional face-to-face interactions, because it does not require extensive coordination and sidesteps many of the pitfalls potentially accompanying group discussions (e.g., groupthink, production blocking[53–57]). Moreover, diagnosticians would benefit from the different interpretations of the raw diagnostic findings (e.g., chest X-ray) without invoking higher information search costs because diagnostic information concerning the patient's history, physical examination, and further diagnostic tests is commonly available to all health professionals involved (through patient charts or electronic health records).

We found that for small groups of 2 or 3 diagnosticians, confidence and (to a lesser extent) seniority could be exploited to increase diagnostic performance in an ill-defined context such as emergency medicine because of their positive relationship with accuracy. Both are usually readily available in practice. Beyond a group size of 4, the follow-the-plurality rule outperformed the other rules, corroborating earlier findings of substantial gains in accuracy when this rule is applied in well-defined domains such as breast and skin diagnostics.[26,28] In practice, the rules may be exploited differently for patients seen in the resuscitation or trauma room (where

many more than 2 physicians are often present and the follow-the-plurality rule can thus be applied) and those seen in the remainder of the ER (where pairs of physicians may base their decisions on either seniority or confidence). We also investigated the performance of very large group sizes of up to 25 as benchmarks to study how well small groups already approximate the collective intelligence of much larger groups. Interestingly, although the marginal improvements diminished at these large group sizes,[58] we still found a continuous improvement for the follow-the-plurality rule. This contrasts with earlier work on combining independent decisions in medical diagnostics that found that collective gains leveled off at much lower group sizes of around 10 diagnosticians.[23,26,28] This difference could be because these studies investigated binary decision tasks, in which there is only 1 wrong decision per case and thus all wrong diagnosticians make the same incorrect decision. In contrast, here we investigated decision scenarios with 20 possible decision outcomes, possibly allowing for higher independence of errors, which could improve the scope for collective intelligence.

The reason why very large group sizes in our study eventually reached near-perfect diagnoses using the follow-the-plurality rule is because in all 6 cases, the correct diagnosis received the largest support among the full sample of 285 participants (Supplemental Figure S1), and the follow-the-plurality rule necessarily amplifies the predominant individual opinion as group size increases.[59,60] Whenever the diagnosis receiving the largest support among all decision makers in the population is the correct diagnosis, then the diagnosis based on all those decision makers is necessarily the correct diagnosis by virtue of the mathematical definition of the follow-the-plurality rule (or, equivalently, the multivariate hypergeometric distribution[61] when the number of draws equals the population size). The speed with which the follow-the-plurality rule approached perfect accuracy as group size increased depends, however, on the specific empirical distribution of answers (e.g., the margin with which the correct diagnosis is preferred over the second most frequent diagnosis; Supplemental Figure S1). Note that for "wicked cases,"[47,48,62] where an incorrect diagnosis receives the largest support in the population, the plurality rule necessarily converges to 0% accuracy as group size increases. Whether, and to what extent, an incorrect diagnosis receives the largest support in clinical practice is an important empirical question.

One of the underlying mechanisms driving collective improvements is observer variation (a.k.a., interrater agreement).[63,64] Collective improvements can arise only when raters make somewhat different judgments.[21,51,65,66] Previous work has shown that collective gains are highest when pooling decisions of diagnosticians who have low interrater agreement (i.e., low kappa values) as compared to pooling decisions of diagnosticians who have high interrater agreement (i.e., high kappa values), keeping average individual performance constant.[67] Given that we observed only 6 decisions for each participant, we could not directly test the effect of interrater agreement on collective performance, but future studies could investigate the effects of interrater agreement on collective intelligence in emergency medicine. Furthermore, it would be interesting to know what cues different participants use, how they translate the raw diagnostic findings into a cue value where necessary (e.g., from an electroencephalogram [EEG] curve to the judgment of whether an EEG abnormality is present), and how exactly the diversity of cue use and interpretation affects collective gains.[16,68]

Here, we studied the potential benefit of algorithmically combining independent diagnoses by 2 or more (student) physicians of high-fidelity simulated patients. There are, however, a variety of other ways to harness collective intelligence. One is direct interaction during or after the examination of the patient, followed by a joint group decision to resolve disagreements.[48] Previous research has shown that interacting pairs can also substantially outperform the average individual.[35] Future studies should directly compare the available methods to understand under what conditions each approach (e.g., algorithmic combining v. direct interaction) results in performance gains.

Considering limitations, our results are based on experimental and not field data. Although the design of the experiments[35,36] captured several key characteristics of the ill-defined task environment of emergency medicine, including time pressure, incomplete information, and a large set of diagnostic tests and final diagnoses to choose from, the situation in real ERs is even more complicated with many more possible outcomes and uncertainties. Furthermore, we combined decisions made by medical students and not experienced clinicians. As already mentioned above, the benefits of the collective intelligence rules we found were roughly invariant across case difficulty, expertise level, and 2 complementary accuracy criteria. This is consistent with the conjecture that the collective intelligence

benefits we observed would also be found for even more difficult cases, more experienced diagnosticians, and other criteria to evaluate the usefulness of the final diagnoses. However, future research is needed to directly investigate this conjecture and whether our results can be replicated in clinicians working on actual clinical ER cases. In conclusion, combining independent decisions may substantially improve the quality of diagnoses in emergency medicine and reduce inadequate treatments and may thus enhance patient safety.

ACKNOWLEDGMENTS

Data are available on the open-science-framework archive (https://osf.io/73n98). The authors thank Susannah Goss and Anita Todd for editing the manuscript, and Dr. rer. medic. Stefanie Hautz and all members of the Medical Decision Making Group at the Max Planck Institute for Human Development for valuable feedback on earlier versions of this manuscript.

REFERENCES

1. Makary, MA, Daniel, M. Medical error—the third leading cause of death in the US. *Bmj*. 2016;353:i2139.

2. Elstein AS. Clinical reasoning in medicine. In: Higgs J. Clinical Reasoning in the Health Professions. Oxford (UK): Butterworth-Heinemann Ltd.; 1995.

3. Kohn LT, Corrigan JM, Donaldson MS. To Err Is Human: Building a Safer Health System. Washington (DC): Institute of Medicine; 1999.

4. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. N Engl J Med. 1991;324(6):377–84.

5. Lu TC, Tsai CL, Lee CC, Ko PC, Yen ZS, Yuan A, et al. Preventable deaths in patients admitted from emergency department. Emerg Med J. 2006;23(6):452–5.

6. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005;165(13):1493–9.

7. Gandhi TK, Kachalia A, Thomas EJ, Puopolo AL, Yoon C, Brennan TA, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. Ann Intern Med. 2006;145(7):488–96.

8. Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. JAMA. 2009;301(10):1060–2.

9. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med. 2007;49(2):196–205.

10. Kuhn GJ. Diagnostic errors. Acad Emerg Med. 2002;9(7):740–50.

11. Norman G. Research in clinical reasoning: past history and current trends. Med Educ. 2005;39(4):418–27.

12. Balogh EP, Miller BT, Ball JR. Improving Diagnosis in Health Care. Washington (DC): National Academies Press; 2016.

13. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 2008;121(5):S2–23.

14. Surowiecki J. The Wisdom of Crowds. New York: Random House; 2004.

15. Krause S, James R, Faria JJ, Ruxton GD, Krause J. Swarm intelligence in humans: diversity can trump ability. Anim Behav. 2011;81(5):941–8.

16. Page SE. Making the difference: applying a logic of diversity. Acad Manag Perspect. 2007;21(4):6–20.

17. Galton F. Vox populi. Nature. 1907;75(7):450–1.

18. Bruce RS. Group judgments in the fields of lifted weights and visual discrimination. J Psychol. 1935;1(1):117–21.

19. Lorge I, Fox D, Davitz J, Brenner M. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. Psychol Bull. 1958;55(6):337–72.

20. Gordon K. Group judgments in the field of lifted weights. J Exp Psychol. 1924;7(5):389–400.

21. Hogarth RM. A note on aggregating opinions. Organ Behav Hum Perform. 1978;21(1):40–6.

22. Yaniv I. The benefit of additional opinions. Curr Dir Psychol Sci. 2004;13(2):75–8.

23. Kattan MW, O'Rourke C, Yu C, Chagin K. The wisdom of crowds of doctors: their average predictions outperform their individual ones. Med Decis Making. 2016;36:536–40.

24. King AJ, Gehl RW, Grossman D, Jensen JD. Skin self-examinations and visual identification of atypical nevi: comparing individual and crowdsourcing approaches. Cancer Epidemiol. 2013;37(6):979–84.

25. Winkler RL, Poses RM. Evaluating and combining physicians' probabilities of survival in an intensive care unit. Manage Sci. 1993;39(12):1526–43.

26. Wolf M, Krause J, Carney PA, Bogart A, Kurvers RH. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS ONE. 2015;10(8): e0134269.

27. Poses RM, Bekes C, Winkler RL, Scott WE, Copare FJ. Are two (inexperienced) heads better than one (experienced) head? Averaging house officers' prognostic judgments for critically ill patients. Arch Intern Med. 1990;150(9):1874–8.

28. Kurvers RHJM, Krause J, Argenziano G, Zalaudek I, Wolf M. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. JAMA Dermatol. 2015;151:1346–53.

29. Ilgen JS, Humbert AJ, Kuhn G, Hansen ML, Norman GR, Eva KW, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Acad Emerg Med. 2012;19(12):1454–61.

30. Brown TW, McCarthy ML, Kelen GD, Levy F. An epidemiologic study of closed emergency department malpractice claims in a national database of physician malpractice insurers. Acad Emerg Med. 2010;17(5):553–60.

31. Cooper S, Cant R, Connell C, Sims L, Porter JE, Symmons M, et al. Measuring teamwork performance: validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. Resuscitation. 2016;101:97–101.

32. Tan TXZ, Quek NXE, Koh ZX, Nadkarni N, Singaram K, Ho AF, et al. The effect of availability of manpower on trauma resuscitation times in a tertiary academic hospital. PLoS ONE. 2016;11(5):e0154595.

33. Kelleher DC, Kovler ML, Waterhouse LJ, Carter EA, Burd RS. Factors affecting team size and task performance in pediatric trauma resuscitation. Pediatr Emerg Care. 2014;30(4):248–53.

34. Egberink RE, Otten HJ, IJzerman MJ, van Vugt AB, Doggen CJ. Trauma team activation varies across Dutch emergency departments: a national survey. Scand J Trauma Resusc Emerg Med. 2015;23(1):100.

35. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. JAMA. 2015;313(3):303–4.

36. Kunina-Habenicht O, Hautz WE, Knigge M, Spies C, Ahlers O. Assessing clinical reasoning (ASCLIRE): instrument development and validation. Adv Health Sci Educ Theory Pract. 2015;20(5):1205–24.

37. Green SM, Martinez-Rumayor A, Gregory SA, Baggish AL, O'Donoghue ML, Green J, et al. Clinical uncertainty, diagnostic accuracy, and outcomes in emergency department patients presenting with dyspnea. Arch Intern Med. 2008;168(7):741–8.

38. Messick S. Validity. In: Linn R, ed. Educational Measurement. 3rd ed. New York: Macmillan Publishing Co., Inc.; 1989.

39. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7.

40. Yudkowsky R, Otaki J, Lowenstein T, Riddle J, Nishigori H, Bordage G. A hypothesis-driven physical examination learning and assessment procedure for medical students: initial validity evidence. Med Educ. 2009;43(8):729–40.

41. Schuwirth LW, van der Vleuten CP. The use of progress testing. Perspect Med Educ. 2012;1(1):24–30.

42. Ericsson KA, Charness N, Feltovich PJ, Hoffman RR. The Cambridge Handbook of Expertise and Expert Performance. Cambridge (UK): Cambridge University Press; 2006.

43. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, eds. The Nature of Expertise. Hillsdale (NJ): Lawrence Erlbaum Associates, Inc; 1988. p 311-342.

44. Bohle Carbonell K, Stalmeijer RE, Könings KD, Segers M, van Merriënboer JJ. How experts deal with novel situations: a review of adaptive expertise. Educ Res Rev. 2014;12:14-29.

45. Nouns Z, Schauber S, Witt C, Kingreen H, Schüttpelz-Brauns K. Development of knowledge in basic sciences: a comparison of two medical curricula. Med Educ. 2012;46(12):1206–14.

46. Hastie R, Kameda T. The robust beauty of majority rules in group decisions. Psychol Rev. 2005;112(2):494–508.

47. Koriat A. When are two heads better than one and why? Science. 2012;336(6079):360–2.

48. Koriat A. When two heads are better than one and when they can be worse: the amplification hypothesis. J Exp Psychol Gen. 2015;144(5):934–50.

49. Bang D, Fusaroli R, Tylén K, et al. Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. Conscious Cogn. 2014;26:13–23.

50. Yetton PW, Bottger PC. Individual versus group problem solving: an empirical test of a best-member strategy. Organ Behav Hum Perform. 1982;29(3):307–21.

51. Mannes AE, Soll JB, Larrick RP. The wisdom of select crowds. J Pers Soc Psychol. 2014;107(2):276–99.

52. Burroughs TE, Waterman AD, Gallagher TH, Waterman B, Adams D, Jeffe DB, et al. Patient concerns about medical errors in emergency departments. Acad Emerg Med. 2005;12(1):57–64.

53. Janis IL. Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes. Boston: Houghton Mifflin; 1972.

54. Baron RS. So right it's wrong: groupthink and the ubiquitous nature of polarized group decision making. Adv Exp Soc Psychol. 2005;37:219–53.

55. Stroebe W, Nijstad BA, Rietzschel EF. Beyond productivity loss in brainstorming groups: The evolution of a question. Adv Exp Soc Psychol. 2010;43:157–203.

56. Kaba A, Wishart I, Fraser K, Coderre S, McLaughlin K. Are we at risk of groupthink in our approach to teamwork interventions in health care? Med Educ. 2016;50(4):400–8.

57. Madigosky W, Schaik S. Context matters: groupthink and outcomes of health care teams. Med Educ. 2016;50(4):387–9.

58. Yetton P, Bottger P. The relationships among group size, member ability, social decision schemes, and performance. Organ Behav Hum Perform. 1983;32(2):145–59.

59. Condorcet M. Essai sur l'application de l'analyse à la probabilité des décisions redues á la pluralité des voix. Paris: Imprimerie Royale; 1785.

60. Grofman B, Owen G, Feld SL. Thirteen theorems in search of the truth. Theory Decis. 1983;15(3):261–78.

61. Tideman TN, Plassmann F. Developing the aggregate empirical side of computational social choice. Ann Math Artif Intel. 2013;68:31–64.

62. Hertwig R. Tapping into the wisdom of the crowd—with confidence. Science. 2012;336(6079):303–4.

63. LeBreton JM, Senter JL. Answers to 20 questions about interrater reliability and interrater agreement. Organ Res Methods. 2007;11(4):815–52.

64. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.

65. Broomell SB, Budescu DV. Why are experts correlated? Decomposing correlations between judges. Psychometrika. 2009; 74(3):531–53.

66. Clemen RT, Winkler RL. Limits for the precision and value of information from dependent sources. Oper Res. 1985;33(2):427–42.

67. Kurvers RH, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A. 2016;113(31):8777–82.

68. Budescu DV, Yu HT. Aggregation of opinions based on correlated cues and advisors. J Behav Decis Mak. 2007;20(2):153–77.