Erschienen in: JAMA : Journal of the American Medical Association ; 313 (2015), 3. - S. 303-304

Letters

RESEARCH LETTER

Diagnostic Performance by Medical Students Working Individually or in Teams

Diagnostic errors contribute substantially to preventable medical error.[1] Cognitive error is among the leading causes and mostly results from faulty data synthesis.[2] Furthermore, reflecting on their confidence does not prevent physicians from committing diagnostic errors.[1] Diagnostic decisions usually are not made by individual physicians working alone. Our aim was to investigate the effect of working in pairs as opposed to alone on diagnostic performance.

Methods | Volunteer fourth-year medical students recruited via mailing lists at Charité Medical School, Berlin, Germany, participated in the study during June 2013 and gave written informed consent. Their main task was to evaluate 6 simulated cases of respiratory distress on a computer, which were previously validated with students and experts.[3] Participants were randomized (stratified by sex) to work individually or in pairs. Participants received a software demonstration prior to randomization; postrandomization and prior to starting the case assessments, they received a training case.

The 6 diagnostic performance cases were presented in random order. Each case started with a video presentation of a prototypical patient. Thereafter, participants could select, in any order, from 30 diagnostic tests as many as desired, but were instructed to be as fast and accurate as possible. Results were presented as real-world clinical data (eg, auscultation sounds or x-ray images). To complete a case, participants had to select 1 of 20 diagnoses and indicate their confidence.

Dependent variables were diagnostic accuracy (correct or incorrect), number and relevance of diagnostic tests (obtained from expert data[3]), time to diagnoses, time tests would

Table. Accuracy, Background Knowledge, Information Search Measures, and Confidence of Medical Students, Across Cases[a]

	Mean (95% CI)		Test Statistics				
	Individuals	Pairs	*t* Score[b]	*d* Value	*F* Score[b]	η_p^2	*P* Value
Accuracy							
No. of correct cases, mean (median) [IQR]	3.00 (3) [2-4]	4.07 (4) [3-5]	$t_{56} = -2.98$	0.78			.004
Percentage	50.00 (40.53 to 59.47)	67.78 (59.95 to 75.60)					
Background knowledge measured by No. of correct items out of 25, %	75.15 (70.19 to 80.12)	73.26 (69.98 to 76.55)	$t_{86} = 0.65$	0.14			.52
Information Search Measures							
Tests selected							
No.	15.41 (14.57 to 16.24)	15.02 (14.21 to 15.82)			$F_{1,5} = 1.36$[c]	0.21	.30
Relevance[d]	59.83 (58.41 to 61.25)	62.26 (60.89 to 63.63)			$F_{1,5} = 16.74$[c]	0.77	.01
When correct	61.11 (57.93 to 64.29)	61.71 (58.53 to 64.89)			$F_{1,50} = 0.86$[e]	0.02	.36
When incorrect	59.54 (56.75 to 62.32)	64.95 (62.16 to 67.73)			$F_{1,50} = 7.23$[f]	0.13	.01
Time, min:sec							
To diagnosis	2:25 (2:07 to 2:42)	4:27 (4:10 to 4:44)			$F_{1,5} = 44.07$[c]	0.90	.001
For tests in reality	37:26 (33:14 to 41:38)	31:11 (27:08 to 35:14)			$F_{1,5} = 8.42$[c]	0.63	.03
Confidence[g]							
Overall	5.92 (5.59 to 6.26)	7.02 (6.70 to 7.35)			$F_{1,5} = 26.13$[c]	0.84	.004
When correct	6.24 (5.51 to 6.97)	7.53 (6.80 to 8.25)			$F_{1,50} = 19.03$[e]	0.28	< .001
When incorrect	5.28 (4.68 to 5.89)	6.14 (5.53 to 6.74)			$F_{1,50} = 0.66$[f]	0.01	.42
Absolute difference		1.24 (1.05 to 1.44)					
When correct		1.16 (0.88 to 1.44)	$t_{25} = 2.57$	−0.52			.02
When incorrect		1.79 (1.33 to 2.24)					

Abbreviation: IQR, interquartile range.

[a] The topics of the 6 diagnostic performance cases were pneumonia, chronic obstructive pulmonary disease, intoxication, pulmonary edema, pulmonary artery embolism, and unstable ventricular tachycardia.

[b] Results of simple and paired *t* tests and mixed-effects analyses of variance with (1) cases being entered as a random factor and group (individuals vs pairs) as a fixed factor and (2) accuracy and condition as fixed factors.

[c] Group (individuals vs pairs) as a fixed factor.

[d] Defined as case-specific proportion of 20 medical experts who selected each test during test instrument validation[3] (thus ranging in an acquisition rate of 0%-100% of the experts). The relevance indices of all tests selected per case by the student participants was then individually averaged across tests and cases.

[e] Accuracy as a fixed factor.

[f] Accuracy and group (individuals vs pairs) as fixed factors.

[g] Indicated on a Likert scale (1 = least to 10 = most confident).

take in reality, and confidence (on a Likert scale from 1 = least to 10 = most confident). Before the main task, participants took a multiple-choice test about respiratory diseases to check whether knowledge about the topic differed between groups (individual vs pairs).

A required sample size of 117 was determined, assuming the pairs would correctly diagnose 1 more case ($\alpha = 0.05$, $\beta = 0.2$, dropout = 5%). The study design was approved by the Charité Medical School institutional review board. We conducted *t* tests for confidence (within pairs), participant characteristics, accuracy, and relevant knowledge (between conditions), and analyses of variance for all other analyses in SPSS version 21 (SPSS Inc) with a 2-sided significance level of $P < .05$.

Results | Of 88 students recruited, 28 worked individually and 60 in pairs. Participant characteristics did not differ between groups. Pairs were more accurate than individuals (67.78% vs 50.00%; difference, 17.78% [95% CI, 5.83%-29.73%]; $P = .004$) despite having comparable knowledge about the topic and selecting an equal number of diagnostic tests (**Table**). Pairs selected more relevant tests on average, but did so only when incorrect.

Pairs needed 2:02 minutes (95% CI, 1:37 to 2:28 minutes) longer than individuals to reach a diagnosis, but their selected tests would have taken 6:15 minutes (95% CI, −12:08 to −0:21 minutes) less in reality. Pairs were more confident than individuals, but their confidence was not better calibrated (same difference between correct and incorrect cases). Within pairs, confidence between participants differed more when incorrect than when correct (1.79 vs 1.16; difference, 0.63 [95% CI, 0.12 to 1.13]; $P = .02$).

In addition, to assess whether pairs might perform better because they are statistically more likely to contain a knowledgeable member,[4] we randomly paired all participants of the individual group into 28 simulated pairs and used the performance of the more confident member as this pair's performance. The procedure was repeated 1000 times and performance averaged. The accuracy of simulated pairs was comparable with individuals (mean, 56.73%; 95% CI, 49.72%-63.74%) but below that of real pairs ($F_{2,83} = 6.75$, $\eta_p^2 = 0.14$, $P = .002$).

Discussion | Working collaboratively reduced diagnostic errors among medical students. As in previous research,[2] neither differences in knowledge nor in amount and relevance of acquired information explained the superior accuracy of the pairs; neither did the statistically increased likelihood of containing a knowledgeable member. Similar to other studies,[4] collaboration may have helped correct errors, fill knowledge gaps, and counteract reasoning flaws.

Pairs were more confident in diagnoses overall; future studies should examine whether a difference in confidence between members could indicate incorrect diagnoses and thus further reduce diagnostic error, as results suggest.

Limitations are the sample of participants (senior students, not physicians) and the test procedure (simulated, not real patients). In addition, all information was shared, which may be different in real clinical settings.[5]

Wolf E. Hautz, MD, MME
Juliane E. Kämmer, PhD
Stefan K. Schauber
Claudia D. Spies, MD
Wolfgang Gaissmaier, PhD

Author Affiliations: Department of Anesthesiology and Intensive Care Medicine, Charité Campus Mitte and Campus Virchow Klinikum, Berlin, Germany (Hautz, Spies); Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany (Kämmer); Institute of Medical Sociology and Rehabilitation Science, Charité Universitätsmedizin Berlin, Berlin, Germany (Schauber); Department of Psychology, University of Konstanz, Konstanz, Germany (Gaissmaier).

Corresponding Author: Juliane E. Kämmer, PhD, Max Planck Institute for Human Development, Center for Adaptive Rationality, Lentzeallee 94, 14195 Berlin, Germany (kaemmer@mpib-berlin.mpg.de).

Author Contributions: Drs Kämmer and Schauber had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Hautz and Kämmer contributed equally.
Study concept and design: Hautz, Kämmer, Spies, Gaissmaier.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Hautz, Kämmer, Schauber.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Kämmer, Schauber.
Obtained funding: Kämmer, Spies.
Administrative, technical, or material support: Hautz, Kämmer, Schauber, Spies.
Study supervision: Hautz, Spies, Gaissmaier.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Spies reported receiving grants from Ethical Committee Vienna Faculty of Medicine, Zon-Mw-Dutch Research Community, Care Fusion, Deltex, Fresenius, Hutchinson, Medizinische Congressorganisation Nürnberg, Novartis, Pajunk, Grünenthal, Köhler Chemie, Roche, Orion Pharma, Outcome Europe Sàrl, University Hospital Stavanger, Arbeitsgemeinschaft Industrieller Forschungsvereinigungen, Bund Deutscher Anästhesisten, Bundesministerium für Bildung und Forschung, Deutsche Krebshilfe, Deutsches Zentrum für Luftund Raumfahrt, German Research Society, Gesellschaft für Internationale Zusammenarbeit, Inner University Grants, Stifterverband, and the European Commission; and receiving personal fees from B. Braun Foundation, ConvaTec International Service GmbH, Pfizer Pharma, Vifor Pharma, Fresenius Kabi, and Georg Thieme Verlag. No other disclosures were reported.

Funding/Support: This study was supported by grants from the Ministry of Education, Youth and Sciences of Berlin awarded to Dr Spies.

Role of the Funder/Sponsor: The Ministry of Education, Youth and Sciences of Berlin had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We acknowledge Fabian Stroben (Charité Universitätsmedizin, Berlin, Germany) for his support in data acquisition; Olga Kunina-Habenicht, PhD (University of Frankfurt/Main, Frankfurt, Germany), Olaf Ahlers, MD (Charité Universitätsmedizin), and Michel Knigge, PhD (University of Halle, Halle, Germany), for their contribution to the development of test cases; Olga Kunina-Habenicht, PhD (University of Frankfurt/Main), and Raimund Senf, MD (Charité Universitätsmedizin), for their support in acquiring expert test data; and Stefanie Hautz (Charité Universitätsmedizin) for her critique of the manuscript. None received financial or other compensation for their contributions.

1. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med*. 2008;121(5)(suppl):S2-S23.

2. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Med Educ*. 2010; 44(1):94-100.

3. Blaum W, Kunina-Habenicht O, Spies C, et al. TEmE: a new computer-based test of the development of medical decision making competency in students [in German]. http://www.egms.de/static/en/meetings/gma2010/10gma064.shtml. Accessibility verified December 10, 2014.

4. Laughlin PR, VanderStoep SW, Hollingshead AB. Collective vs individual induction. *J Pers Soc Psychol*. 1991;61(1):50-67.

5. Christensen C, Larson JR Jr, Abbott A, et al. Decision making of clinical teams. *Med Decis Making*. 2000;20(1):45-50.