Peer Discussion Decreases Practice Intensity and Increases Certainty in Clinical Decision-Making Among Internal Medicine Residents

Neha Bansal Etherington, MD
Caitlin Clancy, MD
R. Benson Jones, MD
C. Jessica Dine, MD, MHSP
Gretchen Diemer, MD

ABSTRACT

Background Team-based decision-making has been shown to reduce diagnostic error, increase clinical certainty, and decrease adverse events.

Objective This study aimed to assess the effect of peer discussion on resident practice intensity (PI) and clinical certainty (CC).

Methods A vignette-based instrument was adapted to measure PI, defined as the likelihood of ordering additional diagnostic tests, consultations or empiric treatment, and CC. Internal medicine residents at 7 programs in the Philadelphia area from April 2018 to June 2019 were eligible for inclusion in the study. Participants formed groups and completed each item of the instrument individually and as a group with time for peer discussion in between individual and group responses. Predicted group PI and CC scores were compared with measured group PI and CC scores, respectively, using paired *t* testing.

Results Sixty-nine groups participated in the study (response rate 34%, average group size 2.88). The measured group PI score (2.29, SD = 0.23) was significantly lower than the predicted group PI score (2.33, SD = 0.22) with a mean difference of 0.04 (SD = 0.10; 95% CI 0.02–0.07; $P = .0002$). The measured group CC score (0.493, SD = 0.164) was significantly higher than the predicted group CC score (0.475, SD = 0.136) with a mean difference of 0.018 (SD = 0.073; 95% CI 0.0006–0.0356; $P = .022$).

Conclusions In this multicenter study of resident PI, peer discussion reduced PI and increased CC more than would be expected from averaging group members' individual scores.

Introduction

Physician decision-making has been studied in a variety of settings in attempts to identify cognitive biases, increase diagnostic accuracy, and decrease resource utilization.[1,2] However, as medical complexity has increased, clinicians now rarely operate in isolation, and clinical decisions are increasingly made within formal or informal clinical teams. The National Academy of Medicine has emphasized physician-led team-based care as a tool to improve patient-centered care, transitions of care, and health outcomes.[3] To our knowledge, few studies have attempted to simulate a team-based approach to care in assessing decision-making among practicing physicians or physicians-in-training.

Team-based approaches to decision-making have been shown to have a positive impact through aggregation of individual judgments and team-based discussion. Aggregation of individual judgments of dermatologists independently reviewing skin lesions and radiologists independently reviewing mammograms was found to be more accurate in the detection of skin and breast cancer than individual judgments alone.[4–6] On a larger scale, the Human Diagnosis Project demonstrates that pooling clinician assessments through an online platform improves diagnostic accuracy across clinical disciplines.[7] These studies focus on what has been termed "collective intelligence," the aggregation or pooling of individual clinical assessments entered asynchronously by clinicians who are temporally and/or geographically separated. Few studies have attempted to capture the more complex dynamics that arise among individuals working together in teams. Medical students working in pairs were less prone to diagnostic errors and more confident in their answers than medical students working individually, and systematic peer cross-checking among emergency medicine physicians is associated with a decreased rate of adverse events.[8,9] On the other hand, group decision-makers are vulnerable to an additional set of biases not applicable to the individual and fall into predictable decision-making traps. Biases introduced

DOI: http://dx.doi.org/10.4300/JGME-D-20-00948.1

Editor's Note: The online version of this article contains the study instrument of clinical vignettes.

into group decision-making such as group polarization and escalation of commitment have potential implications for patient care and patient safety.[10,11]

Previous studies have shown that the learning environment a resident trains in explains variation in practice intensity (PI) more than individual demographic characteristics or personality traits.[12] For example, residents who are supervised by attendings who are more likely to prescribe brand name statins are more likely to prescribe brand name statins themselves.[13] It is important to understand the ways in which learning environments influence resident development of practice patterns in order to better identify modifiable targets for improvement of graduate medical education. One way in which residents hone their clinical decision-making skills is through discussion with their peers.

The purpose of this study is to assess the impact of peer discussion on internal medicine (IM) resident PI, defined as the likelihood of ordering diagnostic tests, empiric therapies, or subspecialty consultations. We aim to assess the effect of peer discussion on PI and clinical certainty (CC). As a secondary analysis, we aim to identify resident characteristics associated with greater response to peer discussion and likelihood of compromise.

Objectives
The purpose of this study was to assess the effect of peer discussion on resident practice intensity (PI) and clinical certainty (CC).

Findings
Peer discussion is associated with a decrease in practice intensity and an increase in clinical certainty.

Limitations
This was a vignette-based study that may not reflect real world clinical practice.

Bottom Line
This study examines one way in which a team-based approach may affect clinical decision-making.

Methods

The study was conducted from April 2018 to June 2019 at 7 IM residency programs ranging from small community-based programs to large university programs in the Philadelphia area (Abington Hospital, Christiana Care Health System, Hahnemann University Hospital, Main Line Health System, Temple University Hospital, Thomas Jefferson University Hospital, and the University of Pennsylvania Health System). All IM residents including preliminary, transitional, and categorical residents were included in the study.

We used an instrument that has been previously described in the literature to measure PI and CC and adapted it to capture individual and group responses.[12] The instrument consists of 34 brief clinical vignettes, designed to describe situations in which ordering a diagnostic test, empiric treatment, or subspecialty consultation may or may not be indicated (provided as online supplementary data). After a description of the clinical scenario, participants were asked whether they would order a particular diagnostic test, empiric treatment, or subspecialty consultation. Each item has 4 possible responses: "Definitely no," "Probably no," "Probably yes," and "Definitely yes," which are assigned a numerical value of 1 to 4. A PI score was calculated, defined as the mean of all vignette responses. A CC score was calculated, defined as the proportion of "Definitely no" and "Definitely yes" answers. In addition, we collected demographic information including age, gender, training program, and level of training.

The instrument was administered in person by 3 authors (N.B.E, C.C., R.B.J.), in place of or following regular didactic sessions at times and locations that were determined to be convenient by individual residency program leadership. A chance at a $45 cash gift card was offered as an incentive for participation.

Participants were asked to form groups of 3 (although groups of 2–5 were allowed), comprised of either interns or second- and third-year IM residents. The instrument was accessed via an online survey platform (REDCap, Vanderbilt University, Nashville, TN) on the participant's handheld device. Participants were directed to complete the instrument by individually reading and selecting an answer to each vignette, conferring with their group, and then selecting a group answer to the same vignette. Participants were instructed that their group answer should be a consensus.

To examine the effect of group work on PI, we calculated a predicted PI score for each group, which we defined as the numerical mean of the individual PI scores of group members. We then used paired *t* testing, with the hypothesis that the difference between the predicted group PI score and the measured group PI score would be greater than zero, or in other words, that working in groups would lower PI more than expected from the numerical mean of the group's individual PI scores. Similarly, we used paired *t* testing to compare predicted group CC score with the measured group CC score, hypothesizing that the difference between the predicted group CC score and the measured group CC score would be less than zero, or that working in groups would increase CC more than expected from the mean individual CC scores of the group members.

As a secondary analysis, we explored the factors influencing whether a participant made a high number of answer switches after conferring with their group. We calculated the total number of answer switches, defined as moving from "Definitely no" or "Probably no" to "Definitely yes" or "Probably yes," or vice versa. We defined high-switch individuals as participants who switched answers more than 1 SD above the mean number of switches. We then created a logistic regression model with high-switch as the dependent variable and gender, intern status, individual PI score, and individual CC score as the independent variables. We hypothesized that women, interns, and those with higher individual PI scores would be more likely to have a high number of switches, and that those with a high individual CC score would be less likely to have a high number of switches.

A P value of $< .05$ was used to determine statistical significance. This study was reviewed and deemed exempt by the Institutional Review Board at Thomas Jefferson University Hospital and was either reviewed or deferred by each individual study site.

Results

Of the 651 IM residents eligible for inclusion in the study, 222 participated (34% response rate). They formed a total of 69 groups with an average group size of 2.88 (median = 3, range 2–5, SD = 0.74). Baseline individual and group characteristics are reported in TABLE 1.

The mean individual PI score was 2.33 with an SD of 0.26, where a higher score (maximum 4) indicates a more intense practice style. The mean individual CC score was 0.49 with an SD of 0.17, where a higher score (maximum 1) implies that the respondent was more certain in their answers (TABLE 2).

PI in Groups

The average predicted group PI score was 2.33 (SD = 0.22, 95% CI 2.28–2.38), and the average measured group PI score was 2.29 (SD = 0.23, 95% CI 2.23–2.34). We found that the measured group PI score was significantly lower than the predicted group PI score, with a mean difference of 0.04 (SD = 0.10; 95% CI 0.02–0.07; $P = .0002$; Cohen's d = 0.443; TABLE 1).

CC in Groups

The average predicted group CC score was 0.475 (SD = 0.136, 95% CI 0.443–0.508), and the average measured group CC score was 0.493 (SD = 0.164, 95% CI 0.454–0.533). We found that the measured group CC score was significantly higher than the predicted group CC score, with a mean difference of 0.018 (SD = 0.073; 95% CI 0.0006–0.0356; $P = .022$; Cohen's d = 0.248; TABLE 2).

TABLE 1
Study Population

Individual Characteristics (N = 222)	n (%)
Age	Mean 28.9 (median 28, range 24–38)
PGY	
PGY-1	96 (43)
PGY-2	52 (23)
PGY-3	52 (23)
PGY-4	7 (3)
Unknown	15 (7)
Gender	
Male	114 (51)
Female	91 (41)
Prefer not to say	5 (3)
Unknown	12 (5.4)
Group Characteristics (N = 69)	**n (%)**
Group size	Mean 2.88 (median 3, range 2–5)
Gender composition	
Mixed gender	44 (64)
Single gender	23 (33)
Unknown	2 (3)

Abbreviation: PGY, postgraduate year.

Answer Switching in Groups

The mean number of switches, or changes from yes to no or vice versa from the individual answer to the group answer, was 2.86 (median = 3; SD = 2.29; range

TABLE 2
Primary Results

Variable	Mean (SD)	95% CI
Individual practice intensity (PI)	2.33 (0.26)	. . .
Predicted group PI	2.33 (0.22)	2.28–2.38
Measured group PI	2.29 (0.23)	2.23–2.34
Mean difference (measured-predicted) in group PI	-0.04 (0.10)	-0.02, -0.07
Individual clinical certainty (CC)	0.49 (0.17)	
Predicted group CC	0.475 (0.136)	0.443–0.508
Measured group CC	0.493 (0.164)	0.454–0.53
Mean difference (measured-predicted) group CC	0.018 (0.073)	0.0006–0.0356

TABLE 3
Logistic Regression Analysis for High-Switch Status

Variable	OR	95% CI	*P* Value
Female gender	2.26	0.908–5.63	.08
Intern	3.42	1.26–9.29	.016
Individual practice intensity	7.49	1.12–49.9	.038
Individual clinical certainty	0.022	0.001–0.470	.014

0–12). After dichotomizing the outcome to define high-switch individuals (> 5 switches, > 1 SD from the mean), we performed logistic regression with gender, intern, individual PI score, and individual CC score as the independent variables. Female gender was associated with an OR of 2.26 for a high number of switches; however, the term was not significant (P = .08, 95% CI 0.908–5.63). Intern status (OR = 3.42; P = .016; 95% CI 1.26–9.29) and higher individual PI scores (OR = 7.49; P = .038; 95% CI 1.12–49.9) increased the likelihood of having a high number of switches. On the other hand, higher individual CC scores decreased the likelihood of having a high number of switches (OR = 0.022; P = .014; 95% CI 0.001–0.470; TABLE 3).

Discussion

In this multicenter, vignette-based study of resident PI, we found that working in peer groups reduced PI and increased CC more than would be expected from averaging group members' individual scores. These findings suggest that coming to an answer as a group was not a matter of compromising to the middle, but instead that discussing the cases moved the whole group toward less intense practice. This has potential implications for resource utilization and patient outcomes in the clinical setting, particularly when higher intensity practice correlates with lower value care. We also found that peer groups were more certain in their answers. Coupled with prior studies suggesting that group medical decisions outperform those of individuals, this finding is important given that uncertainty is a known driver of low-value care.[14–16] The calculated effect sizes in this study were small; however, in thinking about how to identify meaningful differences in practice pattern from responses to clinical vignettes for which the answer choices were "Definitely yes," "Probably yes," "Probably no," and "Definitely no," we speculated that a change in response from "Probably yes" to "Definitely yes" or vice versa might reflect a change in certainty but that a change in response from "Yes" to "No" or vice versa might reflect a meaningful change in practice. Focusing in on potential changes in practice is what inspired our secondary analysis into answer switching. Unsurprisingly, individuals with higher individual CC scores were less likely to have a high number of answer switches.

Interestingly, the individual PI scores from our study were notably lower than what was seen in the study by Dine and colleagues in 2015 (2.33 vs 2.52).[12] There were differences in participating programs and administration that preclude direct comparison; however, this does suggest a trend toward lower PI in the 4 years since the prior study. It is possible that trainees today are more sensitive to resource utilization and health care costs than they were 4 years ago.

Our study has several limitations. It was limited to one geographic area; however, we included both academic and community-based programs that varied in program size. The previous study evaluating PI in residents showed that residents in a similar group of programs had a range of PI, increasing generalizability. The residents reviewed each case, chose their individual answer, and then discussed it as a group before proceeding to the next case. While the cases were extremely brief, and there was no lag time in between individual response and group discussion for each case, it is possible that exposure to the case a second time influenced the group response. Most importantly, the vignette-based nature may not reflect real world practice patterns. Participants were allowed to choose their own groups, which does not approximate team structure on the wards. Furthermore, participants anticipating sharing their answers with their peers may have led to social desirability bias influencing individual and group responses.

The results of this study open up many potential avenues for future research. This study used an instrument that measures PI and CC. It is not designed to measure the appropriateness of more or less intense practice, which is an important question worthy of further investigation. We did not try to standardize the demographic composition of groups. Future work should identify if there are group factors that contribute to group dynamics and influence group outcomes. It would be very interesting to explore both qualitatively and quantitatively the factors that arise in group decision-making that influence group outcomes. We speculate that an individual's initial response may be based on reflexive type I reasoning and that the intervention of peer discussion may force type II reasoning. Qualitative studies of the discussion resulting in group outcomes could explore if one group member presents new information to the group such as clinical experience or knowledge of the literature that influences the group outcome. Alternatively, given the finding that individuals with lower individual CC switched

answers more frequently, it is possible that peer discussion mitigates uncertainty that the individual would have otherwise alleviated by choosing to order a test or call a consultant. Quantitative research could identify if there are characteristics of individual group members who are more likely to influence the group outcome such as gender or seniority and if there are other drivers of high answer switching. Our study focused only on IM residents; however, future research could expand upon this to include other learners and assess interprofessional collaboration. Finally, while we attempted to approximate clinical decision-making using clinical vignettes, there is of course no substitute for an analysis of group decision-making in actual practice.

Conclusions

In conclusion, peer discussion is associated with decreased PI and increased CC. IM residents with high individual PI were more likely to compromise after peer discussion.

References

1. Hodgson NR, Saghafian S, Mi L, et al. Are testers also admitters? Comparing emergency physician resource utilization and admitting practices. *Am J Emerg Med*. 2018;36(10):1865–1869. doi:10.1016/j.ajem.2018.07.041
2. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med*. 2005;165(13):1493–1499. doi:10.1001/archinte.165.13.1493
3. National Academy of Medicine. Mitchell P, Wynia M, Golden R, et al. Core principles & values of effective team-based health care. https://nam.edu/perspectives-2012-core-principles-values-of-effective-team-based-health-care/. Accessed March 12, 2021.
4. Kurvers RH, Krause J, Argenziano G, Zalaudek I, Wolf M. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol*. 2015;151(12):1346–1353. doi:10.1001/jamadermatol.2015.3149
5. Kurvers RH, Herzog SM, Hertwig R, et al. Boosting medical diagnostics by pooling independent judgments. *Proc Natl Acad Sci U S A*. 2016;113(31):8777–8182. doi:10.1073/pnas.1601827113
6. 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. doi:10.1371/journal.pone.0134269
7. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Netw Open*. 2019;2(3):e190096. doi:https://doi.org/10.1001/jamanetworkopen.2019.0096
8. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. *JAMA*. 2015;313(3):303–304. doi:10.1001/jama.2014.15770
9. Freund Y, Goulet H, Leblanc J, et al. Effect of systematic physician cross-checking on reducing adverse events in the emergency department: the CHARMED cluster randomized trial. *JAMA Intern Med*. 2018;178(6):812–819. doi:10.1001/jamainternmed.2018.0607
10. Mannion R, Thompson C. Systematic biases in group decision-making: implications for patient safety. *Int J Qual Health Care*. 2014;26(6):606–612. doi:10.1093/intqhc/mzu083
11. Christensen C, Larson JR Jr, Abbott A, Ardolino A, Franz T, Pfeiffer C. Decision making of clinical teams: communication patterns and diagnostic error. *Med Decis Making*. 2000;20(1):45–50. doi:10.1177/0272989X0002000106
12. Dine CJ, Bellini LM, Diemer G, et al. Assessing correlations of physicians' practice intensity and certainty during residency training. *J Grad Med Educ*. 2015;7(4):603–609. doi:10.4300/JGME-D-15-00092.1
13. Ryskina KL, Dine CJ, Kim EJ, Bishop TF, Epstein AJ. Effect of attending practice style on generic medication prescribing by residents in the clinic setting: an observational study. *J Gen Intern Med*. 2015;30(9):1286–1293. doi:10.1007/s11606-015-3323-5
14. 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(4):536–540. doi:10.1177/0272989X15581615
15. Kämmer JE, Hautz WE, Herzog SM, Kunina-Habenicht O, Kurvers RHJM. The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. *Med Decis Making*. 2017;37(6):715–724. doi:10.1177/0272989X17696998
16. Allison JJ, Kiefe CI, Cook EF, Gerrity MS, Orav EJ, Centor R. The association of physician attitudes about uncertainty and risk taking with resource use in a Medicare HMO. *Med Decis Making*. 1998;18(3):320–329. doi:10.1177/0272989X9801800310

Neha Bansal Etherington, MD, is Assistant Professor of Clinical Medicine and Director of the Internal Medicine Sub-Internship, Lewis Katz School of Medicine, Temple University, Division of

Hospital Medicine, Temple University Health System; **Caitlin Clancy, MD,** is Instructor of Clinical Medicine, Division of Pulmonary, Allergy and Critical Care, University of Pennsylvania Health System, Perelman School of Medicine, University of Pennsylvania; **R. Benson Jones, MD,** is a Fellow, Division of Endocrinology, Diabetes, and Metabolism, University of Pennsylvania Health System; **C. Jessica Dine, MD, MHSP,** is Associate Professor of Medicine, Division of Pulmonary, Allergy and Critical Care, University of Pennsylvania Health System, and Associate Dean of Faculty Development, Perelman School of Medicine, Leonard Davis Institute of Health Economics, University of Pennsylvania; and **Gretchen Diemer, MD,** is Professor of Medicine, Vice Chair of Education for Medicine, and Senior Associate Dean of Graduate Medical Education and Affiliations, Sidney Kimmel Medical College, Thomas Jefferson University.

Funding: This study was funded by the Northeast Group of Educational Affairs.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank the Northeast Group of Educational Affairs, a subsidiary of the Association of American Medical Colleges.

Corresponding author: Neha Bansal Etherington, MD, Temple University Health System, neha.etherington@tuhs.temple.edu, Twitter @NBEtheringtonMD

Received August 20, 2020; revisions received December 22, 2020, and February 22, 2021; accepted March 1, 2021.