Title

Comparison of Diagnostic Recommendations from Individual Physicians versus the Collective Intelligence of Multiple Physicians in Ambulatory Cases Referred for Specialist Consultation

Journal

Medical Decision Making, 42(3)

ISSN

0272-989X

Authors

Khoong, Elaine C

Nouri, Sarah S

Tuot, Delphine S

et al.

Publication Date

2022-04-01

DOI

10.1177/0272989x211031209

Peer reviewed

HHS Public Access

Author manuscript

Med Decis Making. Author manuscript; available in PMC 2022 April 01.

Published in final edited form as:

Med Decis Making. 2022 April ; 42(3): 293–302. doi:10.1177/0272989X211031209.

Comparison of Diagnostic Recommendations from Individual Physicians versus the Collective Intelligence of Multiple Physicians in Ambulatory Cases Referred for Specialist Consultation

Elaine C Khoong, MD MS[1,2], **Sarah S Nouri, MD MPH**[3], **Delphine Tuot, MD MAS**[2,4,5], **Shantanu Nundy, MD MBA**[6,7], **Valy Fontil, MD MAS MPH**[1,2], **Urmimala Sarkar, MD MPH**[1,2]

[1]Division of General Internal Medicine at Zuckerberg San Francisco General Hospital, Department of Medicine, UCSF, San Francisco, CA

[2]Center for Vulnerable Populations at Zuckerberg San Francisco General Hospital, UCSF, San Francisco, CA

[3]Division of General Internal Medicine, Department of Medicine, UCSF, San Francisco, CA

[4]Division of Nephrology, Department of Medicine, UCSF, San Francisco, CA

[5]Center for Innovation in Access and Quality at Zuckerberg San Francisco General Hospital, UCSF, San Francisco, CA

[6]George Washington University Milken Institute School of Public Health, Washington, DC

[7]Accolade, Inc, Plymouth Meeting, PA

Abstract

Background—Studies report higher diagnostic accuracy using the collective intelligence (CI) of multiple clinicians compared with individual clinicians. However, the diagnostic process is iterative, and unexplored is the value of CI in improving clinical recommendations leading to a final diagnosis.

Methods—To compare the appropriateness of diagnostic recommendations advised by individual physicians versus the CI of physicians, we entered actual consultation requests sent by primary care physicians to specialists onto a web-based CI platform capable of collecting diagnostic recommendations (next steps for care) from multiple physicians. We solicited responses to 35 cases (12 endocrinology, 13 gynecology, 10 neurology) from ≥3 physicians of any specialty through the CI platform, which aggregated responses into a CI output. The primary outcome was appropriateness of individual physician recommendations versus the CI output recommendations,

Corresponding author: Elaine C Khoong, UCSF Box 1364, San Francisco General Hospital, 1001 Potrero Ave, Building 10, Ward 13, San Francisco, CA 94143, elaine.khoong@ucsf.edu, Phone: 628-206-3188, Fax: 628-206-5586.

Prior meeting presentations: Preliminary results were presented at the Society of General Internal Medicine meeting in May 2019.

DECLARATION OF CONFLICTING INTERESTS

Dr. Nundy was previously employed by Human Dx. The remaining authors declare that there is no conflict of interest.

using recommendations agreed upon by two specialists in the same specialty as a gold standard. The secondary outcome was the recommendations' potential for harm.

Results—177 physicians responded. Cases had a median of 7 respondents (interquartile range: 5-10). Diagnostic recommendations in the CI output achieved higher levels of appropriateness (69%) than recommendations from individual physicians (45%) (X^2=5.95, p=0.015). 54% of the CI recommendations vs. 41% of individuals' recommendations were potentially harmful (X^2=2.49, p=0.11).

Limitations—Cases were from a single institution. CI was solicited using a single algorithm/platform.

Conclusions—When seeking specialist guidance, diagnostic recommendations from the CI of multiple physicians are more appropriate than recommendations from most individual physicians, measured against specialist recommendations. Although CI provides useful recommendations, some have potential for harm. Future research should explore how to use CI to improve diagnosis while limiting harm from inappropriate tests/therapies.

INTRODUCTION

Diagnostic errors – defined by the National Academy of Medicine as "the failure to establish an accurate and timely explanation of a patient's health problem or communicate that explanation to the patient"[1] – are frequent and have a significant impact on patient morbidity and mortality.[2–7] Prior studies have shown that five percent of US adults who seek outpatient care annually experience a diagnostic error,[5] and that diagnostic errors may contribute to approximately ten percent of patient deaths.[8,9] Less is known about diagnostic error than other areas of patient safety.[10–12] To advance the field, in 2015, the National Academy of Medicine published *Improving Diagnosis in Health Care*.[13] This report provided eight key suggestions to improve diagnosis and reduce diagnostic errors, including the recommendation to facilitate more effective teamwork among healthcare professionals.

This recommendation has driven patient safety advocates to explore collective intelligence (CI) tools to address diagnostic error and delay. Collective intelligence is a shared intelligence that emerges from a group of individuals acting independently or collectively on the same task. It leverages the fact that a group is likely to outperform an individual in cognitive tasks across a variety of fields.[14,15] Within medicine, CI has shown promise in areas of visual diagnosis (radiology, dermatology),[16,17] where CI has reliably outperformed individual physicians in detecting malignancies. Few studies have explored the benefit of CI for general medical diagnosis, and prior studies in general diagnosis have focused on simulated cases.[18] A study of medical students' abilities to accurately diagnose simulated emergency medicine cases found that the CI outperformed individual students.[19] Another recent study found that the CI of multiple physicians had greater diagnostic accuracy than individual physicians in 1500+ simulated cases written for general practitioners.[20]

The development of digital tools that facilitate collaboration and communication among physicians[21–23] provide new opportunities to investigate and leverage the potential of CI, particularly for physicians who practice in isolated settings. A CI platform open to all

healthcare practitioners is the Human Diagnosis Project (Human Dx), a multinational effort in which physicians and medical students both submit and solve clinical cases, and was the platform used in a prior CI study.[20]

Although CI has shown promise to improve diagnostic accuracy, there is limited literature assessing whether the CI of multiple physicians has utility in the diagnostic process, prior to reaching a definitive final diagnosis. Most literature has focused on diagnostic accuracy and diagnosis as if clinicians reach a diagnosis in a one-step process. However, as noted by the National Academy of Medicine,[1] the diagnostic process is complex and iterative. It involves a repeating cycle of information gathering, information integration and interpretation, and a working diagnosis until the diagnosis is communicated to the patient. Reducing diagnostic errors will require interventions at each stage of the diagnostic process.

In recognition of the multi-step, iterative nature of the diagnostic process in real clinical care, we sought to determine if CI would provide value during earlier stages of the process. Specifically, when a clinician refers a patient to a specialist, the referring clinician is often seeking guidance on next steps for evaluation and care (rather than an immediate diagnosis) – a stage when feedback can most impact diagnostic accuracy.[24] Ideally, all general practitioners would have adequate access to specialty expertise when making decisions outside their clinical expertise, but specialty access is limited in many settings due to time, cost, or availability. Accordingly, our study aimed to: (1) assess the appropriateness of diagnostic steps advised by the CI of multiple physicians versus individual physicians collected on a digital CI platform at an earlier stage of the diagnostic process when cases were referred to a specialist, and (2) describe the potential harm of inaccurate recommendations from the CI output and individual physicians. We hypothesized that the CI output from multiple physicians would provide more appropriate diagnostic recommendations than individual physicians measured against specialist recommendations as the gold standard.

METHODS

Collective Intelligence Platform and Cases

The Human Diagnosis Project (Human Dx) platform is a mobile application that allows individuals with any level of medical training (medical student, resident/fellow, attending physician) to both: 1) submit their own clinical case to elicit feedback on the diagnosis and plan; and 2) contribute feedback on diagnoses and plans for any case submitted by other Human Dx users. For this study, Human Dx users responded by using free text to submit a ranked list of differential diagnoses and suggested next steps for the plan of care. Submitted cases include a one-line summary of the case, a clinical question, and relevant history, physical exam, and diagnostic tests (e.g., laboratory or imaging results). (Appendix 1: Example Case) This study was approved by our institution's institutional review board.

At the time of this study, Human Dx used a 1/n proportionally weighted algorithm based on individual user responses to produce a CI for the case, as previously described [20] (see Appendix 2: Collective Intelligence Rule). In brief, for a submitted case, the Human Dx algorithm creates a CI output composed of respondents' ranked list of diagnoses (collective

differential) and recommendations or next steps (collective plan). (See Appendix 3: Sample Collective Intelligence Output.) This CI output reflects both how frequently a diagnosis or plan appears among all responses and its ranking on each respondent's ordered list (e.g., top diagnosis versus fifth diagnosis), but this automated process does not account for alternative spellings of the same recommendation (e.g., blood pressure measurement or BP measurement).

To acquire a sample of diverse real-life cases to submit to the Human Dx platform, an investigator (EK) reviewed actual specialist consultation requests in endocrinology, gynecology, and neurology cases submitted by clinicians at an integrated healthcare system from 2015 to 2017, using the healthcare system's existing electronic consultation platform (e-consult).[25,26] These three specialties were chosen because they are areas in which primary care clinicians have some knowledge, and within this healthcare system, a specialist is required to review e-consult requests prior to scheduling an in-person specialty clinic appointment. As a result, the specialist consultants often provide recommendations to the referring clinician to advance patient care prior to scheduling a patient for an appointment. For each of the three specialties, an investigator (EK) selected ~15-20 cases for which there were clear diagnostic steps recommended by the e-consult specialist. Most cases were early in the diagnostic process, and no diagnosis was provided through the e-consult communication. We focused on cases where recommendations were provided about next steps for evaluation of the patient (i.e., the plan) rather than diagnosis. Cases were selected to ensure no chief complaint was represented more than twice.

Identification of Specialist-Consensus Recommendations

Few cases had guideline-recommended approaches for working up the patient's complaint. Therefore, to assess the appropriateness of the recommended approach in the study cases, the study team used agreement between two specialist physicians as indicative of a reasonable standard of care and the basis for comparison (i.e., specialist-consensus recommendations), per an established approach drawn from the patient safety literature[27,28] (Figure 1). The first specialist (specialist A in figure 1: step 1) was the initial specialist within the integrated healthcare system that responded to the e-consult. To acquire the recommendations of a second specialist, an investigator (ECK) entered case information onto the Human Dx platform from August 2017 to March 2018. From board-certified specialist users on Human Dx, the study team recruited an endocrinologist, neurologist, and gynecologist to respond to the study cases entered on Human Dx in their specialty (specialist B in figure 1: step 2); each of these specialists responded to the case by submitting a differential diagnosis and planned next steps from September 2017 to April 2018. The Human Dx specialists had access to the exact same information as the other Human Dx users who later responded to the case (Figure 1: step 4). If the e-consult specialist and Human Dx specialist agreed on at least one recommended next step in a submitted case, we included that case in this analysis (N=35). The two specialists reached agreement on at least one recommendation in 12/14 (86%) endocrinology, 13/19 (68%) gynecology, and 10/19 (53%) neurology cases. We designated recommendations that both specialists agreed upon as "specialist-consensus recommendations" and established this as the gold standard against which to assess our outcomes. Each of the 35 cases had one to six specialist-consensus

recommendations. Appendix 4 contains one-line summaries of cases included in this study with their specialist-consensus recommendations.

Data Collection

Collective Intelligence Output—We solicited responses to the 35 cases with specialist-consensus recommendations from Human Dx users from August 2017 to November 2018 until we had a minimum of three respondents, which is the number after which the accuracy of CI plateaus.[20] We only included respondents with medical degrees who practiced within the United States (due to differences in practice patterns and available resources among countries). The physician respondents could be trained in any specialty, including endocrinology, gynecology, or neurology. The CI output for each case was derived from the responses of all physician respondents within the US (excluding the designated specialist B for each case).

Independent Individual Physician Respondents—We designated the first three US-based physician respondents to each case (who could be trained in any specialty) to serve as our comparison cohort of "independent individual physicians" (Figure 1: step 6). Designating these physicians and their individual recommendations was meant to serve as a proxy for a clinician practicing independently in the community without specialty access. The responses from these three respondents were also included in the CI output. There were variable numbers of respondents to each case, and we wished to avoid any one case from contributing more than any other case in comparing independent individual physician respondents' recommendations against the CI output. Accordingly, by capping at three those included in the "independent individual physician" cohort, we ensured that each case had an equal contribution to the outcome.

Participant Characteristics—We collected respondents' level of training, location, and specialty based on self-reported information provided when individuals registered on the Human Dx platform.

Outcomes

Primary outcome—Our primary outcome was the appropriateness of recommended diagnostic next steps, based on agreement with specialist-consensus recommendations. We report this outcome separately for: (a) the CI output, and (b) the first three individual physician respondents (the independent individual physician cohort), which is consistent with the approach used in a prior CI study on the same platform.[20] We defined appropriateness at four different levels (from most to least appropriate):

a. Strict appropriateness: all specialist-consensus recommendations, regardless of the number of recommendations, appear at the top of the ranked list from the CI output or an individual. (If there were five specialist-consensus recommendations, the top five recommendations in the CI output or provided by an individual were the five specialist-consensus recommendations.)

b. Moderate appropriateness: did not meet strict level criteria but all the specialist-consensus recommendations were ranked highly within the CI output or the

individual's recommendations. Specifically, if there were X number of specialist-consensus recommendations, all of them appeared within the top 10 or the top 3 * X number of recommendations (whichever was lower). We used two measurement criteria because the number of recommendations varied across cases from one to six. If a case had only two specialist-consensus recommendations, then both recommendations would need to appear in the top six CI output's recommendations or an individual's recommendations (3 * 2). Alternatively, if a case had five recommendations, we looked only at the top ten, rather than the top 15 (3 * 5) recommendations.

c. Lenient appropriateness: at least one but not all specialist-consensus recommendations appeared within the top 10 or top 3 * X recommendations from the CI output or an individual.

d. Not appropriate: none of the specialist-consensus recommendations appeared within the top 10 or 3 * X recommendations from the CI output or an individual.

Secondary outcome—The secondary outcome was the potential harm of recommendations. We assessed the potential for meaningful harm based on a scale previously employed to classify the harm of errors.[7,29] For each recommendation, we used a binary outcome that focused on potential for at least moderate meaningful harm, which included: initiation or cessation of medications without indication or with contraindications; invasive testing; exposure to unnecessary radiation beyond a plain radiograph; and any other actions determined by two physician investigators (ECK, SSN) to have potential to result in at least moderate harm. We only assessed harm for recommendations that were not specialist-consultant recommendations. For both the CI output recommendations and the independent individual physician cohort recommendations, we report recommendations that were identified as having potential for at least moderate harm that appeared among the top 10 or top 3*X recommendations (whichever was lower, as per the primary outcome).

Analyses

Two investigators (ECK, SSN) independently assessed the primary and secondary outcomes, and manually eliminated duplicate recommendations (e.g., a list of eight recommendations with one set of duplicates ["EKG" and "ECG" each appeared] would be treated as a list of seven recommendations), then reached agreement on appropriateness or harm of all recommendations in all cases. We used descriptive statistics to report characteristics of respondents and all outcomes. We used chi-squared testing to determine differences in the appropriateness and harm of the CI output versus the individual physician cohort. For the CI output, we report both outcomes at a case level (out of 35 total cases). For the independent individual physician cohort, we report the outcomes at the individual level; therefore, the overall assessment of individual physicians is out of 105 physicians (35 cases * 3 physician respondents per case). Our funding source had no role in this study.

RESULTS

Respondent characteristics

A total of 177 physicians responded to the 35 cases (12 endocrinology, 13 gynecology,10 neurology) on Human Dx. A median of 7 physicians (interquartile range [IQR]: 5-10) responded to each case. Table 1 shows characteristics of respondents.

Appropriateness

As shown in figure 2 and detailed in Appendix 5, when combining all levels of appropriateness, the CI output performed better than the independent individual physicians, respectively, in each specialty and overall: endocrinology (7/12) 58% versus (12/36) 33%, gynecology (10/13) 77% versus (18/39) 46%, neurology (7/10) 70% versus (17/30) 57%, and overall (24/35) 69% versus (47/105) 45%. These differences in appropriateness were statistically different for the cases overall (69% vs 45%, $X^2 = 5.95$, p=0.015) but not within any specialty.

Figure 2 displays the level of appropriateness (strict, moderate, or lenient) of the CI output recommendations compared with the independent individual physicians. (Data also shown in Appendix 5.) Unlike the results for any level of appropriateness, the CI output did not consistently perform better than individual physicians when considering only strict appropriateness. CI achieved strict appropriateness for seven cases (20%) overall and for none of the endocrinology cases. In contrast, individual physicians achieved strict appropriateness in 22% of cases overall and at least one of three individual physicians achieved strict appropriateness in each of the three specialties. Among all cases, strict appropriateness was the most common level of appropriateness that individuals achieved (22%) vs moderate (9%) or lenient (14%) appropriateness, whereas the CI achieved higher rates of moderate (23%) and lenient (26%) appropriateness. Some individual physicians achieved strict appropriateness in cases where the CI did not. Specifically, three physicians provided strictly appropriate recommendations for an endocrinology case while the CI did not provide strictly appropriate recommendations for any endocrinology cases.

Harm

When evaluating the top 10 or top 3*X recommendations, one or more recommendations from the CI output for 19 (54%) of the cases (6 endocrinology, 7 gynecology, and 6 neurology; or 54%) had potential for meaningful harm (Table 2). In most cases, one or fewer recommendations in the CI output had potential for harm: endocrinology (median 0.5, IQR 0-2); gynecology (median 1, IQR 0-1); neurology (median 1, IQR 0-2).

Among the individual independent physician cohort, 41% of 105 respondents (43 total: 13 endocrinology; 16 gynecology; 14 neurology) also submitted at least one recommendation with potential for meaningful harm. Of the 105 individuals, the majority provided recommendations with no potential for moderate harm (median 0, IQR 0-1). There were no differences in the number of harmful recommendations recommended by the individual physician cohort when comparing specialties (median 0, IQR 0-1 for all three specialties). These recommendations with potential for harm suggested by 43 individual physicians

were distributed across 80% of the cases (n=28; 8 endocrinology, 11 gynecology, and 9 neurology).

The potential for harm was not statistically significantly different when comparing the CI output in the 35 cases to the recommendations submitted by the 105 physicians in the independent individual physician cohort: 54% (19/35) vs 41% (41/105, $X^2 = 2.49$, p=0.11).

DISCUSSION

Key Findings

In this study, we assessed the performance of the collective intelligence of multiple physicians versus individual physicians in providing appropriate diagnostic steps for a plan of care across three specialties. Although the CI recommendations matched the gold standard specialist recommendations more frequently than most independent individual physicians, the CI recommendations only aligned with specialist recommendations ~70% of the time. Moreover, at least one of the recommendations advised by the CI had potential for at least moderate harm in approximately half of the cases.

Performance of the collective intelligence in literature

Our findings are consistent with prior studies demonstrating that CI outperforms individual physicians in diagnostic accuracy[16,17,19,20] and at a rate of a ~30% improvement in accuracy.[19,20] However, our study expands the literature in an important way. By focusing on actual cases for which a clinician consulted specialists for advice, we provide evidence that CI provides value early in the diagnostic process (such as during the information gathering stage) prior to reaching a definitive final diagnosis. Feedback may be particularly important during these earlier steps in the diagnostic process.[24,30] Specifically, the CI recommendations suggest paths forward in the diagnostic process that may not be explored by a clinician practicing independently. These findings also suggest that CI tools may be beneficial in more complex clinical scenarios, beyond straightforward cases with a known diagnosis.

Despite the potential benefits of a CI tool for these types of e-consult cases, we found that there were instances when individual physicians performed better than the collective. Prior studies have had conflicting results as to whether the CI is better than the best individual physician with comparable training[16] or just better than the average physician with comparable training.[19,20] Absent a methodology to reliably predict if a specific individual is going to perform better than the CI, clinicians may rely on their own perceptions for when to ignore the wisdom of the crowd. Studies have shown that clinicians are overconfident in their diagnostic ability; in particular, clinicians do a poor job of calibrating their diagnostic accuracy in cases with high uncertainty.[31,32] Since cases referred to a specialist are more likely to have higher diagnostic uncertainty, clinicians' overconfidence in these situations may pose a barrier to adoption of CI tools.

Potential harm of the collective intelligence

Although our findings support the potential for CI to improve the diagnostic work-up in cases when clinicians may request advice from a specialist, we did find potential for harm in over half of cases. However, there was a wide range in the type of harm: from inappropriate initiation of prescription medications, to radiation exposure from unnecessary imaging, to invasive diagnostic testing. The harm of an invasive diagnostic test (e.g., diagnostic laparoscopy) is higher than initiating an inappropriate prescription medication (e.g., gabapentin). Our prior studies suggest that clinicians would not blindly follow all recommendations of the CI output.[33] In particular, we previously identified trust in the source of the recommendation as an important consideration that factored into how clinicians would behave after receiving information from a clinical decision support tool.[33]

It is important to also note that even if a collective intelligence recommendation seems relatively benign (additional unnecessary laboratory test), studies increasingly show that overtesting is not only wasteful but can result in patient harm.[34–37] Thus, while collective intelligence helps address some of the most common causes of diagnostic errors, such as failure in hypothesis generation or failure to order a necessary test,[2,38,39] this must be weighed against the harm of pursuing unnecessary tests. Of note, potential harm was also present in nearly half of individuals' recommendations as well, suggesting that the potential harm of CI recommendations may be similar to harm from an inappropriate/inaccurate individual physician care recommendation. Although there was no statistically significant difference in harm between the CI output and independent individual physicians, this issue warrants further exploration in a larger study.

Study limitations

Our study has several important limitations. We used a single CI platform and algorithm. However, prior studies comparing different algorithms used to generate a CI output have demonstrated that algorithmic differences have limited impact on the benefit of CI and that the results of CI collected from one platform are likely generalizable.[16,19,20] The users of Human Dx may not be representative of all physicians, but by removing medical students and international practitioners from our analysis it is more likely these findings are generalizable to practicing primary care clinicians in the US. We collected our real-world patient cases from a single healthcare system and prioritized cases where specialists provided a clear recommendation. This may result in a selection of less complex cases and skew specialist-consensus recommendations to be compatible with local practice patterns, but by requiring that two specialists agreed on a recommendation, we increased the likelihood that recommendations would be considered appropriate in multiple settings. Nonetheless, use of specialist-consensus recommendations as the gold standard has its limitations, since among specialists there will also be disagreement. Despite these limitations, our study adds to the literature by demonstrating the potential for CI to assist primary care clinicians in identifying appropriate evaluative steps during the diagnostic process, not just for determination of a diagnosis.

The path forward

Our findings suggest several areas for further exploration. For clinicians without adequate or timely access to specialty advice, these findings suggest that access to the collective intelligence of multiple clinicians can provide useful recommendations to advance the diagnostic process. This is consistent with prior recommendations that feedback improves diagnosis.[1,24,40,41] Although the collective intelligence tools currently available do not replace the need for timely access to specialty care, they may help improve the diagnostic process in settings with inadequate specialty access.

Given the high rate of inappropriate recommendations, clinicians must be judicious in their acceptance of recommendations. For CI tools, providing users with "quality assurance" data on those providing feedback (e.g., information about expertise, clinical training) is crucial to their acceptance of recommendations.[33] Developers and users of collective intelligence tools should collaboratively explore how to increase the benefits of CI tools (e.g., ensuring necessary diagnostic tests are ordered, expanding the differential diagnosis) while mitigating or providing transparency on how to evaluate the risks from potentially harmful recommendations. Methodologies to help differentiate high quality recommendations from inappropriate recommendations and more highly skilled from less expert contributors may increase the value and uptake of CI tools in actual clinical practice. This is particularly true when considering the growth and improvement in other clinical decision support tools.[42]

This work also provides a pathway toward operationalizing provision of feedback to clinicians. Feedback has been identified as a necessary step to increase diagnostic calibration, a concept that describes when clinicians' confidence in their diagnostic decision making aligns with their actual diagnostic accuracy. Well-calibrated clinicians are better able to identify the correct balance between undertesting (failing to explore a broad enough array of potential diagnoses) and overtesting (exposing patients to the costs and harm of unnecessary tests). [30,32,40,41,43] Tools like the one tested in this study provide an approach for clinicians to acquire real-time feedback on their clinical decision making, which may help facilitate diagnostic improvement. Adoption could be incentivized by providing Continuous Medical Education credits for using these feedback tools.

CONCLUSIONS

The collective intelligence of multiple physicians provides more appropriate recommendations than individual physicians when using board-certified specialist recommendations as a gold standard for next steps in the diagnostic process. This suggests that a CI tool may provide useful evaluation recommendations even before a specialist weighs in, thereby improving timely and accurate diagnoses in settings where access to specialty care might be nonexistent, sparse, or delayed. Recommendations provided by a CI tool should not be blindly followed, as some have potential for meaningful harm. Moreover, clinicians should be wary of higher-risk diagnostic tests/therapies suggested by the collective intelligence. Future work is needed to explore how best to leverage CI (and digital tools to facilitate collaboration) to address gaps in the diagnostic process without exposing patients to additional unnecessary harm. There is also promise in evaluating the use of CI tools to facilitate more timely feedback to clinicians on their medical decision making.

ACKNOWLEDGEMENTS

We would like to acknowledge our Human Dx collaborators who provided us the data for this project as well as the Human Dx users for contributing responses. We would like to also thank our specialist collaborators for contributing a second recommendation on these cases. We would like to thank our collaborators at the Center for Innovation in Access and Quality for providing us with electronic consultation cases. Lastly, we acknowledge Natalie Rivadeneira for creating the graphs for this report and Amy J. Markowitz, J.D. for assisting with manuscript revisions.

Funding

Financial support for this study was provided in part by grants from the following government agencies and foundations. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Moore Foundation (Urmimala Sarkar, Valy Fontil)

Research reported in this publication was supported by the National Heart Lung and Blood Institute of the NIH under Award Number K12HL138046. The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH. (Elaine Khoong)

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the NIH under Award Number KL2TR001870. The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH (Elaine Khoong)

National Institute for Health's National Research Service Award (grant number T32HP19025) (Elaine Khoong and Sarah Nouri)

Blue Shield of California Foundation (Delphine Tuot)

National Cancer Institute (grant number K24CA212294) (Urmimala Sarkar)

APPENDIX 1.: Example case

APPENDIX 2.: Collective intelligence rule

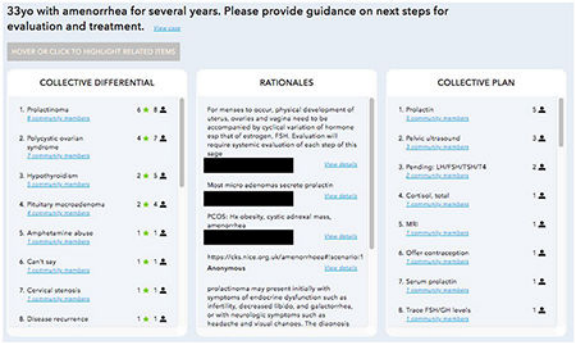
Group size

Collective Intelligence Rule

Physician	Differential Diagnoses					
A	PE	(1/1)	Pneumonia	(1/2)		
B	Pneumonia	(1/1)	COPD	(1/2)	PE	(1/3)
C	PE	(1/1)	Pneumonia	(1/2)	COPD	(1/3)
Collective	PE (1+1/3+1)		Pneumonia (1/2+1+1/2)		COPD (1/2+1/3)	

The collective intelligence is calculated using a weighted formula that considers both the frequency that a specific diagnosis appears on each respondent's list but also where that diagnosis appears on that list. As shown in the example above, pulmonary embolus (PE) is listed as the most likely diagnosis in the collective because it not only appears on all three clinicians' lists, but also overall in the highest position (#1 for physician A and C; #3 for physician B).

APPENDIX 3.: Sample Collective intelligence output



APPENDIX 4: One-line summaries of included cases with specialist-consensus recommendations for each case

Endocrinology (12)

A. 62F many medical problems w recently discovered hypokalemia and adrenal adenoma. Please advise on next steps to determine if patient is adrenalectomy candidate

- Plasma metanephrine
- Dexamethasone suppression
- Renin level in AM
- Aldosterone level in AM

B. 62M HTN, HLD, asthma w persistent hypercalcemia. Please provide advice on further evaluation

- Acquire vitamin 1-25 level

C. 63M HCV/EtOH cirrhosis c/b HCC with hypoglycemia on routine outpatient labs. Please advise on next steps

- Serum fasting glucose

D. 59 F w elevated alk phos x 15 years. Please assist with work-up.

- Vitamin D level

E. 51F w incidental adrenal adenoma. Please assist with work-up

- Plasma metanephrine
- Dexamethasone suppression

F. 39M with headache and low TSH & low FT4. Please assist with next evaluation steps.

- Prolactin

G. 65F primary hyperparathyroidism and thyroid nodule with osteopenia. Please provide advice on next steps

- Sestamibi scan

H. 21F ovarian cysts, irregular menses, elevated DHEA-S. Please advise on next steps

- 17-hydroxyprogesterone level

I. 47M with recent complaint of erectile dysfunction with no ejaculation x 6mo

- AM testosterone
- Prolactin
- LH

J. 49M DM with incidental buffalo hump. Please advise on next steps

- Dexamethasone suppression test

K. 67F osteoporosis and recent compression fx with alendronate contraindications. Please advise on next steps / management / alternative medications.

- Vitamin D level

L. 50M with pituitary macroadenoma. Please advise on appropriate labs to order

- FSH
- Prolactin
- Cortisol

Gynecology (13)

A. 25F G0, h/o ovarian cystic teratoma s/p right oophorectomy (path with mature cystic teratoma) reporting chronic dysmenorrhea. Please advise on next steps in management

- Hormonal treatment

B. 40 yo F with Bartholin cyst. Please advise on next steps

- Sitz bath

C. 25F h/o anovulatory uterine bleeding. Please advise on work-up and management

- Transvaginal ultrasound
- IUD or cyclic provera

D. 63 Korean F s/p hysterectomy 12 years prior on Estradiol. Advise on if, when, and/or how to stop HRT.

- Stop HRT or taper off

E. 27 yo G3P1 @ 27+1 w/ new occurrence of 2 small R labial genital warts. Please advise on next steps.

- TCA or cryotherapy for symptoms

F. 45F w dysuria, hematuria. On CT urogram incidentally found to have adnexal cystic lesion.

- Pelvic ultrasound

G. 42F obese female w h/o unopposed estrogen and inability to conceive. Please provide advice on follow-up EMB and prolactin checks

- Daily provera continuously or progestin IUD

H. 37F w/ recurrent BV. Please advise on next steps

- Suppressive therapy with metronidazole gel twice weekly for 4-6 months

I. 52y G2P2 with hx of adenomyosis/menorrhagia now 1yr post menopause with cervical polyp noted on routine pap. Please advise on next steps.

- Remove polyp

J. 50 y obese F with hypothyroid, HTN, intermittent anemia, h/o irregular & heavy menses. Please advise on next steps.

- IUD placement or hormonal management
- Endometrial biopsy

K. 33yo with amenorrhea for several years. Please provide guidance on next steps for evaluation and treatment.

- Prolactin
- Pelvic ultrasound

L. 40F with abnormal pap. Please provide guidance on timing of repeat pap.

- HPV testing

M. 69 postmenopausal F presenting to new PCP appt w/ c/o intermittent vaginal bleeding s/p previous evaluation. Please provide guidance on further evaluation.

- Pelvic ultrasound or endometrial biopsy with cervical exam

Neurology (10)

A. 62M bipolar and seizure disorder p/w recent “syncopal” episodes. Please advise on next steps before neuro evaluation

- Brain MRI
- EEG
- Dilantin level

B. 40F h/o alcohol use p/w memory complaints x years. Please advise on next evaluation steps.

- Metabolic panel
- LFT
- HIV
- TSH
- B12
- Refer to neuropsych

C. 62M h/o Billroth I p/w bilateral LE neuropathy. Please provide assistance with next steps

- Methylmalonic Acid

D. 50M HCV, opiate dependence p/w worsening bilateral LE peripheral neuropathy. Please advise on next steps to determine etiology

- Hemoglobin A1c
- TSH
- Serum protein electrophoresis (SPEP)

E. 67F w R hand essential tremor x 2 years, worsening. Please advise on next steps

- Propranolol

F. 50M controlled HIV, migraines p/w slurred speech + expressive aphasia a few weeks prior. Please advise on next step

- Start antiplatelet
- Echocardiogram

G. 26 M w left foot drop. Pls advise on next steps.

- Lumbar MRI

H. 40F w worsening migraines x 8-10 years. Please advise on steps prior to neurological evaluation.

- Headache diary

I. 53F HTN, PTSD, h/o BPPV w chronic dizziness. Please assist in next steps to evaluate if dizziness is related to PTSD vs neuro etiology.

- Vestibular physical therapy

J. 63M restless leg symptoms. Please advise on next steps.

- Ropinorole

APPENDIX 5

Appendix 5a:

Appropriateness of Collective Intelligence Recommendations

	Level of Appropriateness			
Specialty	**Strict**	**Moderate**	**Lenient**	**None**
Endocrinology	0 / 12 (0%)	4 / 12 (33%)	3 / 12 (25%)	5 / 12 (42%)
Gynecology	5 / 13 (38%)	3 / 13 (23%)	2 / 13 (15%)	3 / 13 (23%)
Neurology	2 / 10 (20%)	1 / 10 (10%)	4 / 10 (40%)	3 / 10 (30%)
Overall	7 / 35 (20%)	8 / 35 (23%)	9 / 35 (26%)	11 / 35 (31%)

Appendix 5b:

Appropriateness of Individual Physicians

	Level of Appropriateness			
Specialty	**Strict**	**Moderate**	**Lenient**	**None**
Endocrinology	3 / 36 (8%)	4 / 36 (11%)	5 / 36 (14%)	24 / 36 (67%)
Gynecology	14 / 39 (36%)	2 / 39 (5%)	2 / 39 (5%)	21 / 39 (54%)
Neurology	6 / 30 (20%)	3 / 30 (10%)	8 / 30 (27%)	13 / 30 (43%)
Overall	23 / 105 (22%)	9 / 105 (9%)	15 / 105 (14%)	58 / 105 (55%)

When using a binary definition of appropriateness (none vs strict/moderate/lenient), there was a significant difference between individuals vs collective intelligence recommendations among all cases ($X^2 = 5.95$, p=0.015) but not for any specialty: endocrine ($X^2 = 2.35$, p=0.125); gynecology ($X^2 = 3.71$, p=0.054); or neurology ($X^2 = 0.56$, p=0.46).

REFERENCES

1. National Academies of Sciences Engineeering and Medicine. Improving Diagnosis in Health Care [Internet]. Balogh E, Miller B, Ball J, editors. Washington (DC): National Academies Press (US); 2015. Available from: https://www.nap.edu/catalog/21794/improving-diagnosis-in-health-care
2. Schiff GD Diagnostic Error in Medicine. Arch Intern Med [Internet]. American Medical Association; 2009 Nov 9 [cited 2018 May 5];169(20):1881. Available from: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinternmed.2009.333 [PubMed: 19901140]
3. Gandhi, Tejal K, Kachalia A, Thomas EJ, Puopolo AL, Yoon C, Brennan TASD. Missed and Delayed Diagnosis in the Ambulatory Setting: A Study of Closed Malpractice Claims. Ann Intern Med [Internet]. American College of Physicians; 2006 Oct 3 [cited 2018 May 1];147(7):488–96. Available from: http://annals.org/article.aspx?doi=10.7326/0003-4819-145-7-200610030-00006

4. Schiff GD, Kim S, Abrams R, Cosby K, Lambert B, Elstein AS, Hasler S, Krosnjar N, Odwazny R, Wisniewski MF, Mcnutt RA. Diagnosing Diagnosis Errors: Lessons from a Multi-institutional Collaborative Project. In: Henriksen K, Battles JB, Marks ES et al., editor. Adv Patient Saf From Res to Implement (Volume 2 Concepts Methodol [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2005 [cited 2018 May 7]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20492/pdf/Bookshelf_NBK20492.pdf

5. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf [Internet]. 2014 Sep [cited 2018 May 1];23(9):727–731. Available from: http://qualitysafety.bmj.com/content/qhc/23/9/727.full.pdf

6. Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf [Internet]. BMJ Publishing Group Ltd; 2013 Oct 1 [cited 2019 Jul 18];22 Suppl 2(Suppl 2):ii21–ii27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23771902

7. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and Origins of Diagnostic Errors in Primary Care Settings. JAMA Intern Med [Internet]. American Medical Association; 2013 Mar 25 [cited 2019 Jul 23];173(6):418. Available from: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2013.2777 [PubMed: 23440149]

8. Shojania KG, Burton EC, McDonald KM, Goldman L. The autopsy as an outcome and performance measure. Evid Rep Technol Assess (Summ) [Internet]. 2002 Oct [cited 2019 Jul 18];(58):1–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12467146

9. Sonderegger-Iseli K, Burger S, Muntwyler J, Salomon F. Diagnostic errors in three medical eras: a necropsy study. Lancet [Internet]. 2000 Jun 10 [cited 2019 Jul 18];355(9220):2027–2031. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10885353 [PubMed: 10885353]

10. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. BMJ Qual Saf [Internet]. 2015 Feb [cited 2018 Jan 17];24(2):103–110. Available from: http://qualitysafety.bmj.com/content/qhc/24/2/103.full.pdf

11. Al-Mutairi A, Meyer AND, Thomas EJ, Etchegaray JM, Roy KM, Davalos MC, Sheikh S, Singh H. Accuracy of the Safer Dx Instrument to Identify Diagnostic Errors in Primary Care. J Gen Intern Med [Internet]. 2016 Jun 22 [cited 2018 Jan 17];31(6):602–608. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26902245 [PubMed: 26902245]

12. Wachter RM. Why Diagnostic Errors Don't Get Any Respect–And What Can Be Done About Them. Health Aff [Internet]. 2010 Sep [cited 2018 Mar 22];29(9):1605–1610. Available from: https://www.healthaffairs.org/doi/pdf/10.1377/hlthaff.2009.0513

13. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med [Internet]. 2008 May [cited 2019 Oct 9];121(5 Suppl):S2–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18440350

14. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science (80-) [Internet]. 2010 Oct 29 [cited 2019 Apr 16];330(6004):686–688. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20929725

15. Pérez T, Zamora J, Eguíluz VM. Collective Intelligence: Aggregation of Information from Neighbors in a Guessing Game. Marshall JAR, editor. PLoS One [Internet]. 2016 Apr 19 [cited 2019 May 12];11(4):e0153586. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27093274 [PubMed: 27093274]

16. Wolf M, Krause J, Carney PA, Bogart A, Kurvers RHJM. Collective Intelligence Meets Medical Decision-Making: The Collective Outperforms the Best Radiologist. Pavlova MA, editor. PLoS One [Internet]. 2015 Aug 12 [cited 2019 Apr 16];10(8):e0134269. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26267331 [PubMed: 26267331]

17. Kurvers RHJM, Krause J, Argenziano G, Zalaudek I, Wolf M. Detection Accuracy of Collective Intelligence Assessments for Skin Cancer Diagnosis. JAMA Dermatology [Internet]. 2015 Dec 1 [cited 2019 May 12];151(12):1346. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26501400 [PubMed: 26501400]

18. Radcliffe K, Lyson HC, Barr-Walker J, Sarkar U. Collective intelligence in medical decision-making: a systematic scoping review. BMC Med Inform Decis Mak [Internet]. 2019 Dec 9 [cited

2019 Aug 20];19(1):158. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0882-0 [PubMed: 31399099]

19. 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 Mak [Internet]. 2017 Aug 29 [cited 2019 Apr 16];37(6):715–724. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28355975
20. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Netw Open [Internet]. American Medical Association; 2019 Mar 1 [cited 2019 Mar 2];2(3):e190096. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30821822 [PubMed: 30821822]
21. Willman AS. Use of Web 2.0 tools and social media for continuous professional development among primary healthcare practitioners within the Defence Primary Healthcare: a qualitative review. J R Army Med Corps [Internet]. Journal of Medical Internet Research; 2019 Jan 3 [cited 2019 May 12];10(3):jramc-2018-001098. Available from: http://www.jmir.org/2008/3/e22/
22. Boulos MNK, Maramba I, Wheeler S. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. BMC Med Educ [Internet]. BioMed Central; 2006 Dec 15 [cited 2019 May 12];6(1):41. Available from: https://bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-6-41 [PubMed: 16911779]
23. Bacigalupe G Is there a role for social technologies in collaborative healthcare? Fam Syst Heal [Internet]. 2011 Mar [cited 2019 May 12];29(1):1–14. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0022093
24. Kostopoulou O, Rosen A, Round T, Wright E, Douiri A, Delaney B. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. Br J Gen Pract [Internet]. Royal College of General Practitioners; 2015 Jan 1 [cited 2019 Oct 9];65(630):e49–e54. Available from: http://bjgp.org/lookup/doi/10.3399/bjgp15X683161 [PubMed: 25548316]
25. Tuot DS, Murphy EJ, McCulloch CE, Leeds K, Chan E, Chen AH. Leveraging an electronic referral system to build a medical neighborhood. Healthcare [Internet]. 2015 Dec [cited 2019 Apr 15];3(4):202–208. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26699344 [PubMed: 26699344]
26. Chen AH, Murphy EJ, Yee HF. eReferral - A new model for integrated care. N Engl J Med. Massachussetts Medical Society; 2013;368(26):2450–2453. [PubMed: 23802515]
27. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Newhouse JP, Weiler PC, Hiatt HH. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med [Internet]. 1991 Feb 7 [cited 2019 Aug 20];324(6):370–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1987460 [PubMed: 1987460]
28. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, Laffel G, Sweitzer BJ, Shea BF, Hallisey R. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA [Internet]. 1995 Jul 5 [cited 2019 Aug 20];274(1):29–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7791255 [PubMed: 7791255]
29. Singh H, Mani S, Espadas D, Petersen N, Franklin V, Petersen LA. Prescription Errors and Outcomes Related to Inconsistent Information Transmitted Through Computerized Order Entry. Arch Intern Med [Internet]. 2009 May 25 [cited 2019 Jul 23];169(10):982. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2919338/pdf/nihms131429.pdf [PubMed: 19468092]
30. Meyer AND, Singh H. Calibrating how doctors think and seek information to minimise errors in diagnosis. BMJ Quality and Safety. BMJ Publishing Group; 2017. p. 436–438.
31. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. JAMA Intern Med [Internet]. 2013 Nov 25 [cited 2019 May 12];173(21):1952–1961. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23979070 [PubMed: 23979070]
32. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Qual Saf [Internet]. BMJ Publishing Group Ltd; 2019 May 1 [cited 2019 May 12];28(5):352–355. Available from: https://qualitysafety.bmj.com/content/28/5/352

33. Fontil V, Radcliffe K, Lyson HC, Ratanawongsa N, Lyles C, Tuot D, Yuen K, Sarkar U. Testing and improving the acceptability of a web-based platform for collective intelligence to improve diagnostic accuracy in primary care clinics. JAMIA Open [Internet]. 2019 Apr 1 [cited 2019 Feb 4];2(1):40–48. Available from: https://pubmed.ncbi.nlm.nih.gov/31984344/ [PubMed: 31984344]
34. Greenberg J, Green JB. Over-testing: why more is not better. Am J Med [Internet]. Elsevier; 2014 May 1 [cited 2019 May 12];127(5):362–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24269325 [PubMed: 24269325]
35. Moriates C, Soni K, Lai A, Ranji S. The Value in the Evidence. JAMA Intern Med [Internet]. 2013 Feb 25 [cited 2019 May 12];173(4):308. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23358796 [PubMed: 23358796]
36. Vilar-Palop J, Hernandez-Aguado I, Pastor-Valero M, Vilar J, González-Alvarez I, Lumbreras B. Appropriate use of medical imaging in two Spanish public hospitals: a cross-sectional analysis. BMJ Open [Internet]. 2018 Mar 16 [cited 2019 May 12];8(3):e019535. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29549204
37. Wegwarth O, Gigerenzer G. Overdiagnosis and Overtreatment. JAMA Intern Med [Internet]. American Medical Association; 2013 Dec 9 [cited 2019 May 12];173(22):2086. Available from: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2013.10363 [PubMed: 24145597]
38. Neshati H, Sheybani F, Naderi H, Sarvghad M, Soltani AK, Efterkharpoor E, Nooghabi MJ. Diagnostic Errors in Tuberculous Patients: A Multicenter Study from a Developing Country. J Environ Public Health [Internet]. 2018 Nov 13 [cited 2019 May 12];2018:1975931. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30538752 [PubMed: 30538752]
39. Newman-Toker DE. A unified conceptual model for diagnostic errors: underdiagnosis, overdiagnosis, and misdiagnosis. Diagnosis [Internet]. 2014 Jan 1 [cited 2019 May 12];1(1):43–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28367397 [PubMed: 28367397]
40. Meyer AND, Singh H. The Path to Diagnostic Excellence Includes Feedback to Calibrate How Clinicians Think. JAMA [Internet]. American Medical Association; 2019 Feb 26 [cited 2019 Oct 30];321(8):737. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2019.0113 [PubMed: 30735239]
41. Omron R, Kotwal S, Garibaldi BT, Newman-Toker DE. The Diagnostic Performance Feedback "Calibration Gap": Why Clinical Experience Alone Is Not Enough to Prevent Serious Diagnostic Errors. AEM Educ Train. Wiley; 2018 Oct;2(4):339–342. [PubMed: 30386846]
42. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digit Med [Internet]. 2020 Dec 6;3(1):17. Available from: http://www.nature.com/articles/s41746-020-0221-y [PubMed: 32047862]
43. Cifu AS Diagnostic Errors and Diagnostic Calibration. JAMA [Internet]. American Medical Association; 2017 Sep 12 [cited 2019 Oct 30];318(10):905. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.11030 [PubMed: 28828468]

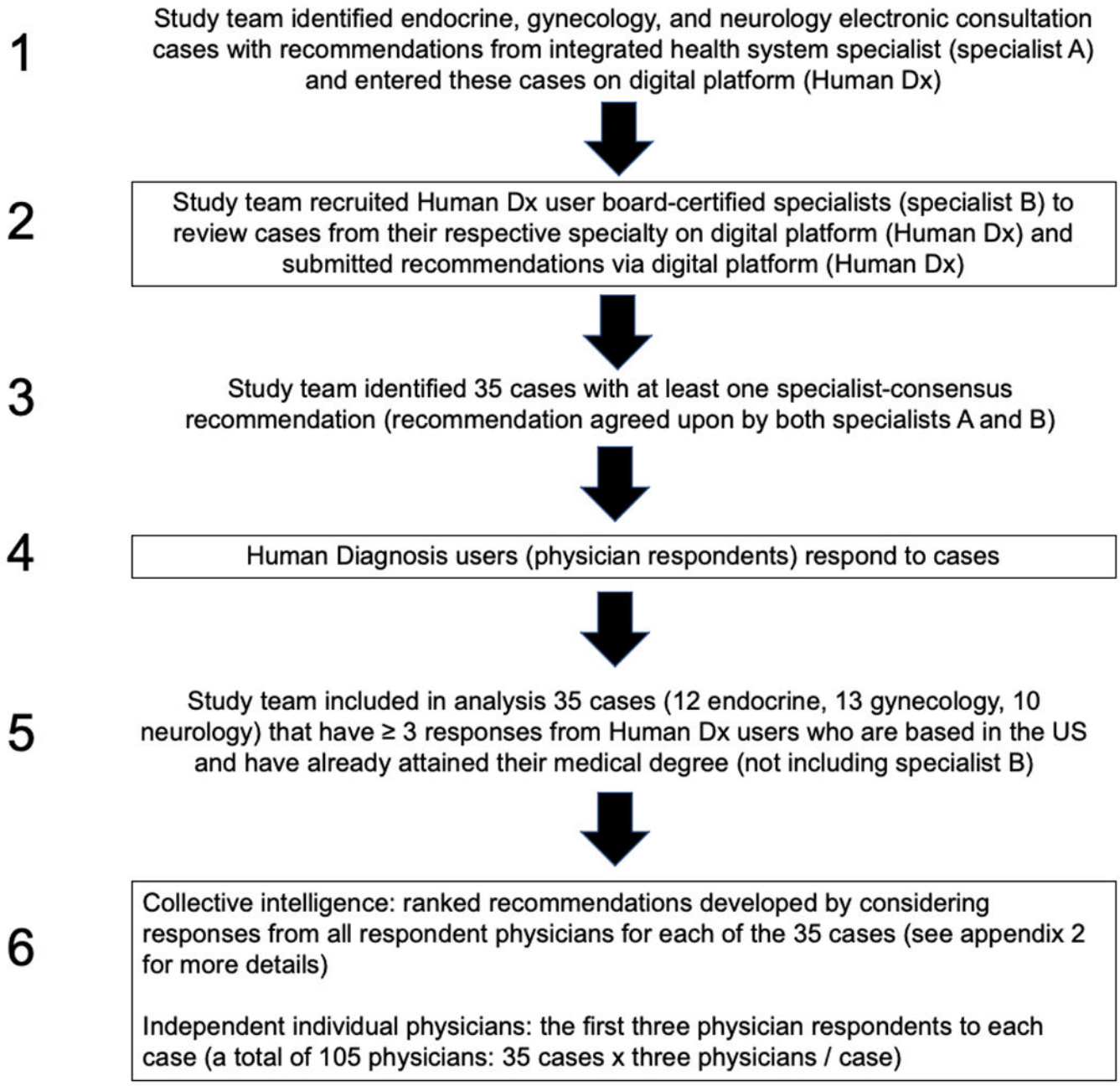


FIGURE 1/.
Study workflow
Note: Specialist A was one of two specialists from each specialty that responded to the integrated health system e-consult. Specialist B was the same clinician for all cases within their specialty and responded to the case on Human Dx.

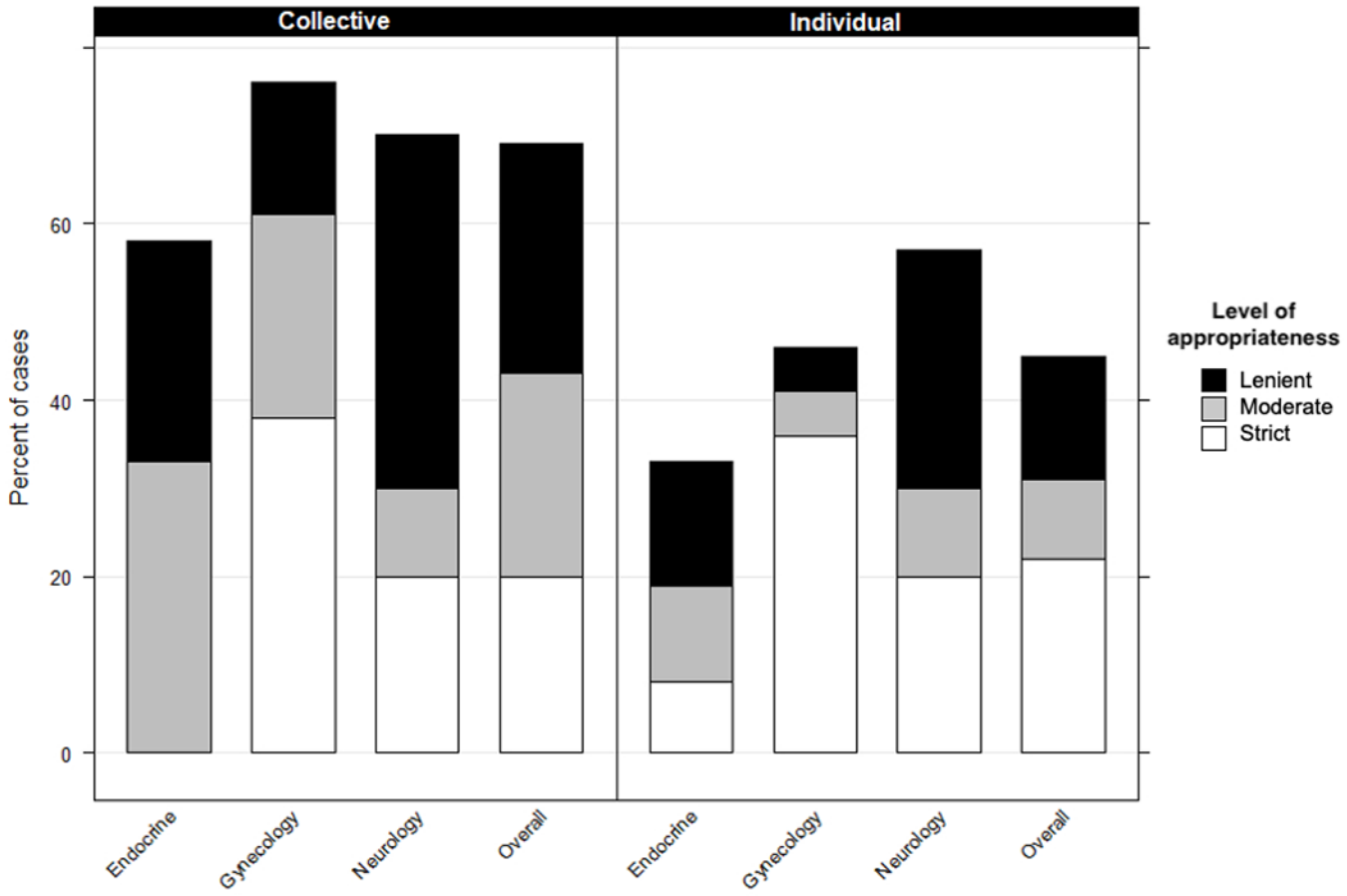


FIGURE 2/.

Appropriateness of recommendations from collective intelligence of multiple physicians vs individual physicians by specialty and in all cases

Table 1

Respondent characteristics

User Characteristic	No. (%) (n = 177)
Training Level	
Attending physician	74 (42%)
Fellow / Resident	103 (58%)
Specialty	
Family Medicine	27 (15%)
Internal Medicine	132 (75%)
General	119 (67%)
Subspecialty	13 (7%)
Other	18 (10%)

Table 2

Collective intelligence recommendations with potential for harm

Case	Recommendations with potential for harm
Endocrinology Case A	Adrenal venous sampling; oral sodium load; surgery
Endocrinology Case C	Computed tomography (CT) of abdomen / pelvis
Endocrinology Case F	Head CT
Endocrinology Case G	Parathyroidectomy; Fine needle aspiration
Endocrinology Case I	Sildenafil; Brain magnetic resonance imaging (MRI)
Endocrinology Case K	Denosumab; Teriparatide injection
Gynecology Case A	Diagnostic laparoscopy
Gynecology Case B	Incision and drainage
Gynecology Case C	Depo-Provera
Gynecology Case H	Biweekly clindamycin; Combined oral contraceptives; oral Flagyl
Gynecology Case I	Biopsy
Gynecology Case J	IV iron infusion
Gynecology Case K	Brain MRI
Neurology Case A	Change antiepileptic; stop phenytoin
Neurology Case D	Nerve conduction studies / electromyography (NCS/EMG); gabapentin
Neurology Case E	Deep brain stimulation
Neurology Case F	Electroencephalogram (EEG)
Neurology Case G	NCS / EMG
Neurology Case H	Triptan; Head CT