Free patient case management tool accurately measures physicians' diagnostic abilities

February 21 2019

Assessing the accuracy and value of an increasingly popular and free online patient management app, researchers at Johns Hopkins Medicine and other institutions say that physicians with more training and experience perform better in selecting appropriate diagnoses for sample patient scenarios.

In a report on the study, published in the Jan. 11 issue of *JAMA Network Open*, the investigators conclude that the app, called The Human Diagnosis Project, or Human Dx, can reliably be used to determine physicians' skills in forming accurate, efficient diagnoses.

"Doctors are constantly trying to stay up to date on current best practices to better provide patients with high-value care," says study co-author Reza Manesh, M.D., assistant program director for clinical reasoning for the Osler Medical Training Program at the Johns Hopkins University School of Medicine, and an assistant professor of medicine. "We do this by reading new research, staying apprised of clinical practice guidelines and attending educational conferences; however, there is no way for doctors to self-assess how well they incorporate that new knowledge into how they make decisions for patients. There is an urgent need to objectively evaluate a physician's clinical reasoning and ensure they are making the right decisions for their patients."

While there is no ideal or single metric to assess diagnostic performance, Manesh adds, "This online tool is a step in the right direction."

Medical "board" exams, for example, feature multiple choice questions to test diagnostic skills, "but if you think about it, patients don't come in to the hospital or the clinic with a list of multiple choice questions," Manesh notes.

Human Dx, free online, combines physician crowdsourcing of expert knowledge with machine learning, merging the collective wisdom of the worldwide medical community for benchmarking and mimicking real-world practice decisions in which clinicians must rule in or rule out dozens of possibilities and symptoms to make the right diagnosis. The tool is designed to facilitate these so-called "differential diagnoses" without the limitations posed by multiple choice.

Manesh and his co-investigators set out to test the ability of Human Dx to assess users' diagnostic skills based on their reported level of experience.

To do that, the investigators analyzed a total of 11,023 Global Morning Report cases solved by U.S. physicians and medical students using Human Dx between Jan. 21, 2016, and Jan. 15, 2017. Cases were solved by 1,738 users (239 attending physicians, 926 resident physicians, 347 intern physicians and 226 medical students) across 170 individual scenarios. The average number of cases solved by each participant was 74. Investigators used three key metrics to assess diagnostic performance : (1) efficiency, a percentile score calculated based on the proportion of findings revealed before the user included the correct diagnosis; (2) accuracy, analyzed by how high on the differential diagnostic scale the correct diagnosis was listed, and by how often the correct diagnosis was listed first; and (3) diagnostic acumen precision performance (DAPP), calculated from a weighted average of the percentiles of both accuracy and efficiency for each attempt to get the right diagnosis.

To use the diagnostic portion of Human Dx, physicians or students log

on to the app to view sample cases that offer basic information about a simulated patient case, as well as X-ray or lab test results. Human Dx provides immediate feedback to users who suggest a differential diagnosis listing potential causes.

Cases cover a range of inpatient and outpatient scenarios from general adult medicine to subspecialty disciplines. All cases and diagnoses offered are first peer-reviewed for accuracy by a member of an independent editorial board comprising attending physicians at academic medical institutions. Johns Hopkins physicians contribute to this peer-review process.

Human Dx has been used by over 16,000 subscribers.

Based on the ranking of the correct diagnosis compared with all users who solved the cases, the highest average score for attending physicians (assumed to be the most highly trained and experienced group overall) was 76.9 (the highest possible score any one user could get was 100). For residents, the highest average score was 76.8, and for interns it was 74.7. Students—the least experienced group—showed a highest average score of 68.8.

Overall, investigators found that attending physicians had higher accuracy scores than medical students (difference of 8.1 percent), residents (difference of 8 percent) and interns (difference of 5.9 percent). Attending physicians also had higher efficiency compared with residents (difference of 4.8 percent), interns (difference of 5 percent) and students (difference of 5.4 percent); and significantly higher DAPP scores than residents (difference of 2.6 percent), interns (difference of 3.6 percent) and students (difference of 6.7 percent).

Forty percent of participants (496 people) were affiliated with one of the U.S. News and World Report top 25 medical schools. DAPP scores were

highest for attending physicians affiliated with one of these top 25 schools compared with attending physicians for other institutions, a difference of 80 versus 72. In addition, residents affiliated with one of these schools had higher DAPP scores compared with their nonaffiliated peers (75 versus 71), as did interns (75 versus 69).

Thirty-two percent of participants (417 people) were affiliated with one of the top 25 institutional recipients of National Institutes of Health research grants. Again, DAPP scores were higher among attending physicians, residents and interns affiliated with these institutions compared with their nonaffiliated peers (81 versus 72 for attending physicians, 75 versus 71 for residents, and 76 versus 69 for interns).

The Institute of Medicine, in its 2015 report "Improving Diagnosis in Health Care," estimated that most people will experience at least one diagnostic error in their lifetime, and that "getting the diagnosis right" is a crucial component of effective health care, Manesh says. It has been estimated that $100 billion or more may be wasted annually in the U.S. as a result of inaccurate diagnosis, Manesh adds. "Diagnostic errors affect an estimated 12 million Americans each year, and likely cause more harm to patients than all other medical errors combined."

The authors caution there were several limitations of the study, including that poor-quality attempts to solve the cases were dropped from the data set, so effort by users likely varied. In addition, unlike actual patient encounters, in which clinical information must be gathered and synthesized over time, the case simulations provide clinical data up front. There also may have been performance bias in favor of more technologically advanced individuals.

Manesh and colleagues also are studying the ability of Human Dx to identify master diagnosticians compared with those with good or average abilities. This could help institutions identify their best diagnosticians.

Human Dx was launched in 2015 as a partnership among public and private organizations involved in medical education and practice, including the Association of American Medical Colleges, the American Medical Association, boards of medical specialties and academic medical centers such as Johns Hopkins, Harvard, the University of California, San Francisco, and MIT. Financial support has come from the MacArthur Foundation, Moore Foundation and others. The Human Diagnosis Project says it has thousands of users in 80 countries and 500 medical institutions involved in building the project, and that ultimately they plan to give any physician, organization or patient direct access to its toolbox.

Provided by Johns Hopkins University School of Medicine