# Letters

RESEARCH LETTER

## Comparison of Physician and Computer Diagnostic Accuracy

The Institute of Medicine recently highlighted that physician diagnostic error is common and information technology may be part of the solution.[1] Given advancements in computer science, computers may be able to independently make accurate clinical diagnoses.[2] While studies have compared computer vs physician performance for reading electrocardiograms,[3] the diagnostic accuracy of computers vs physicians remains unknown. To fill this gap in knowledge, we compared the diagnostic accuracy of physicians with computer algorithms called symptom checkers.

Symptom checkers are websites and apps that help patients with self-diagnosis. After answering a series of questions, the user is given a list of rank-ordered potential diagnoses generated by a computer algorithm. Previously, we evaluated the diagnostic accuracy of 23 symptom checkers using 45 clinical vignettes.[4] The vignettes included the patient's medical history and had no physical examination or test findings. In this study we compared the diagnostic performance of physicians with symptom checkers for those same vignettes using a unique online platform called Human Dx.

**Methods** | Human Dx is a web- and app-based platform on which physicians generate differential diagnoses for clinical vignettes. Since 2015, Human Dx has been used by over 2700 physicians and trainees from 40 countries who have addressed over 100 000 vignettes.

The 45 vignettes, previously developed for the systematic assessment of online symptom checkers,[4] were disseminated by Human Dx between December 2015 and May 2016 to internal medicine, family practice, or pediatrics physicians who did not know which vignettes were part of the research study. There were 15 high, 15 medium, and 15 low-acuity condition vignettes and 26 common and 19 uncommon condition vignettes.[4] Physicians submitted free text ranked differential diagnoses for each case. Each vignette was solved by at least 20 physicians.

Given that physicians provided free text responses, 2 physicians (S.N. and D.M.L.) hand-reviewed the submitted diagnoses and independently decided whether the participant listed the correct diagnosis first or in the top 3 diagnoses. Interrater agreement was high (Cohen κ, 96%), and a third study physician (A.M.) resolved discrepancies (n = 60).

We used $\chi^2$ tests of significance to compare in physicians' performance. Physician diagnosis accuracy was compared with previously reported symptom checker accuracy for these same vignettes using 2-sample tests of proportion.[4] The study was exempt from Harvard's institutional review board and participants were not compensated.

**Results** | Of the 234 physicians who solved at least 1 vignette, 211 (90%) were trained in internal medicine and 121 (52%) were fellows or residents (**Table 1**).

Physicians listed the correct diagnosis first more often across all vignettes compared with symptom checkers (72.1% vs 34.0%, $P < .001$) as well as in the top 3 diagnoses listed (84.3% vs 51.2%, $P < .001$) (**Table 2**).

Table 1. Physician Diagnostic Accuracy, Stratified by Physician Characteristic

| | No. (%) | | Listed, % (95% CI) | | | |
|---|---|---|---|---|---|---|
| Physician Characteristic | Physicians | Vignettes Completed | First | *P* Value[a] | Top 3 | *P* Value[a] |
| All physicians | 234 (100) | 1105 (100.0) | 72.1 (69.5-74.8) | NA | 84.3 (82.2-86.5) | NA |
| Tenure | | | | .97 | | .15 |
| Attending | 71 (30) | 475 (43.0) | 71.8 (67.7-75.9) | | 82.7 (79.3-86.1) | |
| Fellow/resident[b] | 121 (52) | 487 (44.1) | 72.5 (68.5-76.5) | | 84.4 (81.2-87.6) | |
| Intern[c] | 42 (18) | 143 (12.9) | 72.0 (64.6-79.5) | | 89.5 (84.4-94.6) | |
| Specialty | | | | .13 | | .01 |
| Internal medicine | 211 (90) | 937 (84.8) | 73.0 (70.2-75.8) | | 85.5 (83.2-87.7) | |
| Internal medicine subspecialty | 23 (10) | 168 (15.2) | 67.3 (60.1-74.4) | | 78.0 (71.6-84.3) | |
| Total No. of vignettes ever completed using Human Dx[d] | | | | .52 | | .83 |
| 1-30 | 127 (54) | 380 (34.4) | 70.5 (65.9-75.1) | | 84.2 (80.5-87.9) | |
| 31-120 | 65 (28) | 353 (31.9) | 71.7 (66.9-76.4) | | 83.6 (79.7-87.5) | |
| ≥121 | 42 (18) | 372 (33.7) | 74.2 (69.7-78.7) | | 85.2 (81.6-88.8) | |

Abbreviation: NA, not applicable.

Totals may not add up to 100% owing to rounding.

[a] Comparison across all physicians.

[b] Includes postgraduate year 2 and above.

[c] Includes postgraduate year 1 only.

[d] Includes vignettes completed outside of the vignettes disseminated in this study. Categories of usage were defined to evenly distribute the number of vignettes completed by the physicians.

Table 2. Physician and Symptom Checkers' Diagnostic Accuracy, Stratified by the Acuity Level and Prevalence of the Conditions Described by the Clinical Vignettes

| Vignette Characteristic | No. (%) | | Listed, % (95% CI) | | | |
|---|---|---|---|---|---|---|
| | | | First[a] | | Top 3[a] | |
| | Vignettes Completed by Human Dx Physicians | Vignettes Completed by Symptom Checkers | Human Dx Physicians | Symptom Checkers[b] | Human Dx Physicians | Symptom Checkers[a] |
| All vignettes | 1105 (100) | 770 (100) | 72.1 (69.5-74.8) | 34.0 (30.7-37.4) | 84.3 (82.2-86.5) | 51.2 (47.4-54.3) |
| Acuity level[c] | | | | | | |
| High | 398 (36.0) | 263 (34.2) | 79.1 (75.1-83.2) | 24.3 (19.1-29.6) | 89.2 (86.1-92.3) | 39.5 (33.6-45.5) |
| Medium | 376 (34.0) | 260 (33.7) | 70.7 (66.1-75.4) | 37.7 (31.8-43.6) | 84.3 (80.6-88.0) | 56.9 (50.9-63.0) |
| Low | 331 (30.0) | 247 (32.1) | 65.3 (60.1-70.4) | 40.5 (34.3-46.7) | 78.5 (74.1-83.0) | 57.5 (51.3-63.5) |
| Vignette prevalence[d] | | | | | | |
| Common | 639 (57.8) | 457 (59.4) | 69.6 (66.1-73.2) | 38.1 (33.6-42.5) | 83.3 (80.4-86.2) | 55.6 (51.6-60.7) |
| Uncommon | 466 (42.2) | 313 (40.6) | 75.5 (71.6-79.5) | 28.1 (23.1-33.1) | 85.8 (82.7-89.0) | 44.7 (38.4-49.3) |

Totals may not add up to 100% owing to rounding.

[a] $P<.001$ for all comparisons between physicians and symptom checkers.

[b] Results described by Semigran et al.[4] Full version of clinical vignettes available at: http://www.bmj.com/content/bmj/suppl/2015/07/07/bmj.h3480.DC1/semh025489.ww1.pdf.

[c] Acuity level of vignettes defined by Semigran et al.[4] Differences across physicians and across symptom checkers for this category were statistically significant ($P < .001$).

[d] We defined "common" diagnoses as those that accounted for more than 0.3% of ambulatory visits (or >3 764 082 visits) in the United States in 2009 to 2010. These totals were compiled from data gathered by the Centers for Disease Control and Prevention, the National Ambulatory Medical Care Survey, and the National Hospital Ambulatory Medical Care Survey. Differences across physicians and across symptom checkers for this category were statistically significant ($P < .05$) except for the difference between the rate that physicians listed the correct diagnosis in the top 3 for common vs uncommon vignettes.

Across physicians, they were more likely to list the correct diagnosis first for high-acuity vignettes (vs low-acuity vignettes) and for uncommon vignettes (vs common vignettes). In contrast, symptom checkers were more likely to list the correct diagnosis first for low-acuity vignettes and common vignettes (Table 2).

**Discussion** | In what we believe to be the first direct comparison of diagnostic accuracy, physicians vastly outperformed computer algorithms in diagnostic accuracy (84.3% vs 51.2% correct diagnosis in the top 3 listed).[4] Despite physicians' superior performance, they provided the incorrect diagnosis in about 15% of cases, similar to prior estimates (10%-15%) for physician diagnostic error.[5] While in this project we compared diagnostic performance, future work should test whether computer algorithms can augment physician diagnostic accuracy.[6]

Key limitations included our use of clinical vignettes, which likely do not reflect the complexity of real-world patients and did not include physical examination or test results. Physicians who chose to use Human Dx may not be a representative sample of US physicians and therefore may differ in diagnostic accuracy. Symptom checkers are only 1 form of computer diagnostic tools, and other tools may have superior performance.

**Hannah L. Semigran, BA**
**David M. Levine, MD, MA**
**Shantanu Nundy, MD**
**Ateev Mehrotra, MD, MPH**

**Author Affiliations:** Harvard Medical School, Boston, Massachusetts (Semigran, Mehrotra); Brigham and Women's Hospital, Boston, Massachusetts (Levine); The Human Diagnosis Project, Washington, DC (Nundy).

**Corresponding Author:** Ateev Mehrotra, MD, MPH, Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (mehrotra@hcp.med.harvard.edu).

**Published Online:** October 10, 2016. doi:10.1001/jamainternmed.2016.6001

**Author Contributions:** Dr Mehrotra had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
*Concept and design*: All authors.
*Acquisition, analysis, or interpretation of data*: Semigran, Levine, Nundy.
*Drafting of the manuscript*: Semigran.
*Critical revision of the manuscript for important intellectual content*: All authors.
Statistical analysis: Semigran.
*Administrative, technical, or material support*: Levine, Mehrotra.
*Study supervision*: Nundy, Mehrotra.

**Conflict of Interest Disclosures:** Dr Nundy is an equity holder of The Human Diagnosis Project, the creators of Human Dx. No other disclosures are reported.

**1**. The National Academies of Science Engineering and Medicine. *Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press; 2015.

**2**. Topol EJ. The Future of Medicine Is in Your Smartphone. *The Wall Street Journal*; January 9, 2015; The Saturday Essay.

**3**. Poon K, Okin PM, Kligfield P. Diagnostic performance of a computer-based ECG rhythm algorithm. *J Electrocardiol*. 2005;38(3):235-238.

**4**. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*. 2015;351:h3480.

**5**. Graber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf*. 2013;22(suppl 2):ii21-ii27.

**6**. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med*. 2012;27(2):213-219.

LESS IS MORE

## Prediabetes Risk in Adult Americans According to a Risk Test

The Diabetes Prevention Program and other studies found that individuals with impaired glucose tolerance (based on a 75-g oral glucose tolerance test) can decrease their risk of type 2