Common Laboratory Results Frequently Misunderstood by a Sample of Mechanical Turk Users

Nabeel Qureshi[1] Ateev Mehrotra[2,3] Robert S. Rudin[2] Shira H. Fischer[2]

[1] RAND Corporation, Santa Monica, California, United States
[2] RAND Corporation, Boston, Massachusetts, United States
[3] Harvard Medical School, Boston, Massachusetts, United States

Address for correspondence Shira H. Fischer, MD, PhD, RAND Corporation, 20 Park Plaza, Suite 920, Boston, MA 02116, United States (e-mail: sfischer@rand.org).

Appl Clin Inform 2019;10:175–179.

Abstract

Objectives More patients are receiving their test results via patient portals. Given test results are written using medical jargon, there has been concern that patients may misinterpret these results. Using sample colonoscopy and Pap smear results, our objective was to assess how frequently people can identify the correct diagnosis and when a patient should follow up with a provider.

Methods We used Mechanical Turk—a crowdsourcing tool run by Amazon that enables easy and fast gathering of users to perform tasks like answering questions or identifying objects—to survey individuals who were shown six sample test results (three colonoscopy, three Pap smear) ranging in complexity. For each case, respondents answered multiple choice questions on the correct diagnosis and recommended return time.

Results Among the three colonoscopy cases ($n = 642$) and three Pap smear cases ($n = 642$), 63% (95% confidence interval [CI]: 60–67%) and 53% (95% CI: 49–57%) of the respondents chose the correct diagnosis, respectively. For the most complex colonoscopy and Pap smear cases, only 29% (95% CI: 23–35%) and 9% (95% CI: 5–13%) chose the correct diagnosis.

Conclusion People frequently misinterpret colonoscopy and Pap smear test results. Greater emphasis needs to be placed on assisting patients in interpretation.

Keywords
- Amazon Mechanical Turk
- health literacy
- laboratory results

Background and Significance

An increasing number of patients have access to their test results in their medical record through patient portals.[1] Providing access to medical records is a key "meaningful use" criterion[2] used for incentive payments to providers under the new Merit-based Incentive Payment System (MIPS) program for physician payment. While such access is highly valued by patients,[3] prior qualitative work has raised concern that many patients have trouble understanding the information they are receiving.[4] Text-based test results such as pathology, radiology, and procedure reports that are provided in patient portals are written for communication between medical providers; therefore, the reports typically include complex medical language and extraneous information for patients. Ideally, health care providers would interpret these results for their patients via a phone call, letter, or during a visit. However, most patients view this information in the portal first.[5] Poor comprehension may be particularly high for those with low medical literacy.[4] The most basic information a patient can pull from a laboratory report is what the results were and if there is any follow-up needed.[6]

received
August 19, 2018
accepted after revision
January 17, 2019

DOI https://doi.org/10.1055/s-0039-1679960.
ISSN 1869-0327.

Objectives

Our goal was to quantify the extent to which people interpret medical screening tests correctly and whether they can determine, from the resulting report they receive after a test, what their diagnosis is and the recommended time to return to the doctor.

Methods

We focused on the results of colonoscopy and Pap smear for testing because they are common; tens of millions[7,8] of Americans receive these screenings each year, and the results are delivered in text (vs. a numerical test result). Real-world examples of colonoscopy and Pap smear results were reviewed by two physicians (A.M. and S.H.F.) who modified the reports to remove any patient identifiers.

We chose three different laboratory reports of varying complexity for each of the two procedures (see ►Table 1): for each kind of procedure, case 1 was the least complex and case 3 was the most complex. Difficulty was determined qualitatively and comparatively, based on how explicitly the results were stated and how easy the results were to find in the report. The cases are included in ►Supplementary Material S1 and S2 (available in the online version).

Amazon Mechanical Turk (mTurk) is an online labor market that has been used in numerous research studies in behavioral research as well as increasingly in health.[9] We used mTurk to quickly and efficiently field survey questions on health literacy to a large number of individuals.

We recruited a sample of adults (n = 1,000) from mTurk. We began with a preliminary survey to obtain informed consent and allow for randomized assignment of participants to different testing categories as well as to collect information about age, gender, their education level, whether they had any medical training, whether they had insurance, and if they had any difficulty reading, writing, or understanding information. Our sample comprised only those with higher than a 95% approval rating to achieve high data quality.[10] Of those, 642 completed the follow-up survey for inclusion in the study. Each participant then received a random procedure result and associated follow-up pathology of varying complexity for one of the two possible test result types—colonoscopy or Pap smear—using their Mechanical Turk ID number. The preliminary survey also ensured that a single individual did not take multiple surveys and that all responders were United States based and had high (95%) approval ratings to try and ensure the quality of collected data.[10]

All participants were paid federal minimum wage for the time they took to complete the test based on a priori estimates of time to complete. While there is no clear relationship between wage and quality, we paid a higher wage in case it was helpful.[11] The total payout to participants was approximately $485.

The survey was completed in 7 days based on reaching our minimum target number of participants and budget limit. For each case (colonoscopy and Pap smear), the respondent was asked to select the correct diagnosis and also asked when an appropriate time would be to return to the doctor based on this result from preset multiple-choice options (for full set of questions and answer choices, see ►Supplementary Material S3 [available in the online version]).

Results

Among the 642 respondents, women made up the majority (67%) and 46% were between 18 and 34 years of age. The single largest age category was ages 25 to 29 at 19%.

For diagnosis, among all colonoscopy results, 63% of the 642 respondents correctly chose the diagnosis, and for Pap smear results, 53% of the 642 respondents correctly interpreted the diagnosis (see ►Table 2). At baseline, the percent correct diagnosis for colonoscopy ranged from 29% (n = 203)

Table 1 Test diagnosis and return times, by case

Colonoscopy			
Complexity	**Case 1 (low complexity)**	**Case 2 (moderate complexity)**	**Case 3 (high complexity)**
Correct diagnosis	Everything was normal	There was a growth, but it was not related to cancer, and there was swelling on the person's anus or bottom (hemorrhoids).	There was a growth and it was cancer and there were little pouches on the colon wall (diverticulosis).
Correct return time	In 1–2 years	In 3–10 years	In 6 months
Pap smear			
Complexity	**Case 1 (low complexity)**	**Case 2 (moderate complexity)**	**Case 3 (high complexity)**
Correct diagnosis	Everything was normal	The test was not normal, it is not cancer, but it might turn into cancer later, and there was a yeast infection.	Cancer was found
Correct return time	In 3–5 years	In 1 year	Within the next 2 weeks

Table 2 Rates of respondents being able to identify correct diagnosis and return time

Colonoscopy			
	Case 1 ($n = 221$) (low complexity)	Case 2 ($n = 218$) (moderate complexity)	Case 3 ($n = 203$) (high complexity)
Rate of identifying correct diagnosis	93%	65%	29%
Rate of identifying correct return time	18%	89%	26%
Pap smear			
	Case 1 ($n = 208$) (low complexity)	Case 2 ($n = 220$) (moderate complexity)	Case 3 ($n = 214$) (high complexity)
Rate of identifying correct diagnosis	98%	54%	9%
Rate of identifying correct return time	9%	41%	29%

to 93% ($n = 221$) and the percent correct diagnosis for Pap smears ranged from 9% ($n = 214$) to 98% ($n = 208$) (see ►Fig. 1). There was no association between age of respondent and correct choice of diagnosis.

The return time was selected correctly for colonoscopy by 18% for the low complexity case, 89% for the intermediate complexity case, and 26% for the complex case. The correct return time for Pap smears was selected by 9% for the low complexity case, 41% for the intermediate complexity case, and 26% for the complex case.

Discussion

Driven in part by federal incentives, there has been a proliferation of patient portals where patients can view their test results. While such efforts are patient centric and quite popular, prior qualitative studies have suggested that test results are sometimes not understood.[12] The findings of our survey echo these concerns with two common cancer screening tests, colonoscopy and Pap smear results.

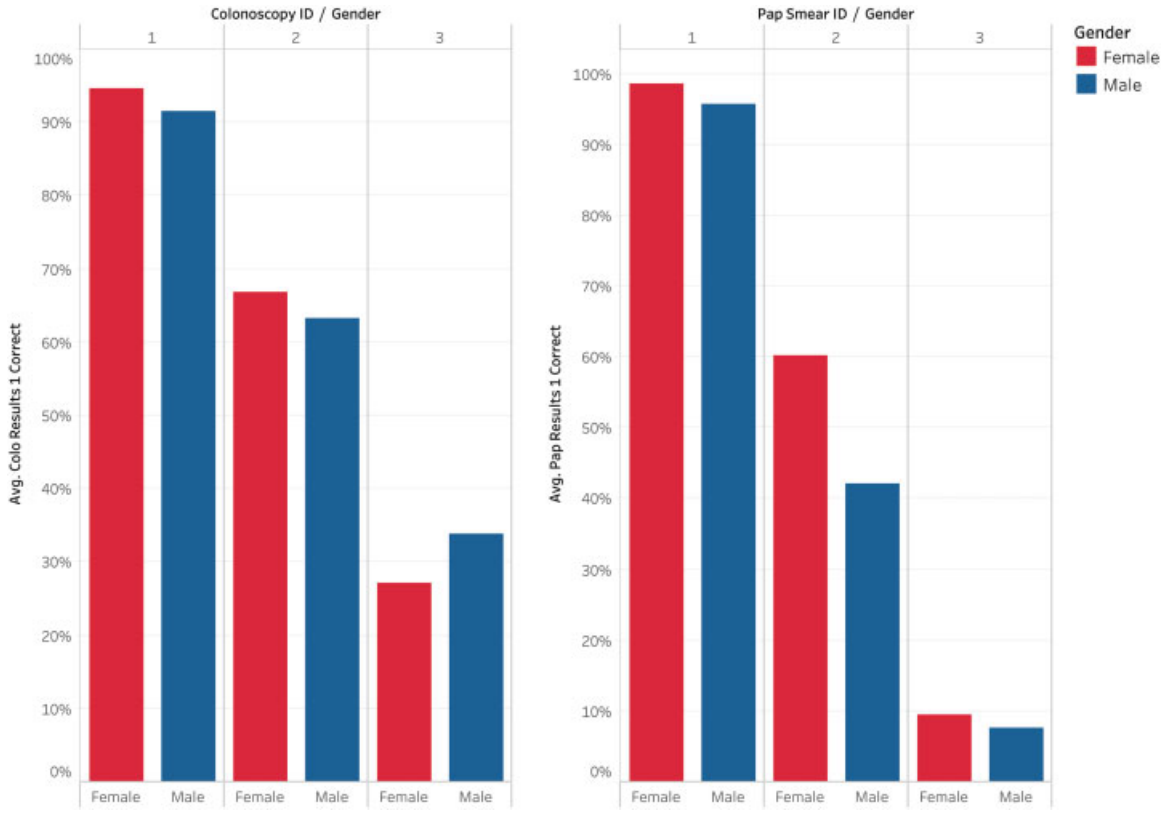

Fig. 1 Percent correct by type of test (colonoscopy versus Pap), case (1 versus 2 versus 3, with 1 being the least complex), and gender. More complex cases (3 versus 2 versus 1) had lower rates of percent correct ($p < 0.001$). Differences by gender were not significant.

Successful interpretation of the results, not surprisingly, did vary by the complexity of the results. When the results were normal, most respondents were able to successfully identify that this was a normal diagnosis. However, when there was a more serious diagnosis such as a cancer, a minority were able to recognize this serious diagnosis. In contrast, identification of when they should follow up did not vary by complexity of the diagnosis. The results indicate that some information in these laboratory reports will be obvious to most patients (i.e., normal results for colonoscopy and Pap smear), while other information will be difficult for nearly all patients (i.e., when to return for a follow-up for normal Pap smear results or when to return for abnormal colonoscopy results), which is important for patients to understand to get the right care at the right time. Given the high education rate of our test population, based on mTurk demographics, the risk of those of lower educational background to misunderstand results would be even higher.

There are several ways that health systems could assist patients in interpretation. The most ideal would be that the ordering physician always writes a note when the test is released to the patient to assist with interpretation. That is the expectation of most health systems. Unfortunately, prior work has highlighted that this provider interpretation occurs in the minority of cases. Providers have many competing demands, and this is a time-intensive task. Possibly technology could be used to assist. For example, natural language processing methods could be used to "translate" these reports from the complex medical language to a language that is at the appropriate reading comprehension level. Such tools could potentially be used to ease the burden on providers or be used in an automated matter. Without better patient understanding, patient portals could stress and confuse patients rather than inform them. We did not design this study to distinguish between positive and negative findings, but since physicians often do not provide follow-up after negative results, while positive results have more serious consequences, follow-up work could include examining whether understanding should focus on positive results or negative results and when.

One novel innovation of this project was the use of the mTurk platform to field this short survey. This builds on other studies that have used mTurk to understand patient comprehension.[13,14] mTurk allowed for a rapid and cost-effective approach (<$500) to easily assess patients' understanding of typical test results without any instruction across a large population of varying demographics. The National Institutes of Health is encouraging researchers to consider mTurk for research.[15] Given the speed and low cost, mTurk's ideal role may be in targeting next steps of a study (i.e., experimenting with different types of information presentation), improve practices (i.e., targeting on patient education materials such as difficult-to-interpret results instead of all results), or to clarify assumptions about health literacy or comprehension. mTurk does have its limitations. mTurk respondents tend to be more educated, younger, and a higher percent female than the average U.S. population. However, it has been argued that it is more representative than typical American college samples, which are commonly used in health literacy studies.[16] Another concern is the quality of the responses from mTurk. Respondents are paid per task and therefore they have a financial incentive to answer questions as quickly as possible. It is possible to screen for higher quality respondents through preliminary testing, or to select them using reputation scores, as we did.[17]

There are other important limitations of this study beyond the use of the mTurk platform. Respondents were not viewing their own test results. Therefore, they were not likely to be in the same frame of mind as those patients reviewing their own cancer screening tests. Their emotional involvement was therefore less and they are likely less anxious and less invested in understanding the results than a real patient. The impact of this bias on comprehension is unclear. mTurk respondents could be less invested to determine the right answer and simply rush to complete the survey. However, if anxiety clouds the understanding of a real patient, our mTurk respondents could make less mistakes. Additionally, there are other limitations to mTurk, including the issue of representation,[18] concern for bots confounding results or intentional fraud,[19,20] concern for low data quality (we did not use attention check questions, though we limited our sample to high reputation workers),[10] and more. Finally, we used a convenience sample and not a probability sample, thus the point estimates may not be what will be seen in the general population,[17] but the relative rates should still be informative.

In summary, our results highlight that it may be common for patients to misinterpret common laboratory results.

Clinical Relevance Statement

Patients often misinterpret common laboratory results. We recommend changes to results' format to include clearer explanation of the results and expected follow-up to ensure accurate comprehension.

Multiple Choice Questions

1. What is the main benefit of using Amazon Turk?
   a. Speed.
   b. Accuracy.
   c. Representativeness.
   d. Validity.

   Correct Answer: The correct answer is option a, speed.

2. Which is a downside of using Amazon Turk?
   a. The sample may not be representative.
   b. The cost is prohibitive.
   c. It takes a long time to get results.
   d. It has not been used for medical studies yet so we cannot trust it.

   Correct Answer: The correct answer is option a, representativeness.

Protection of Human and Animal Subjects

This project was deemed exempt by RAND's institutional review board.

Conflict of Interest

None declared.

References

1 Patel V, Johnson C. Individuals' use of online medical records and technology for health needs. ONC Data Brief 40, 2018. Available at: https://www.healthit.gov/sites/default/files/page/2018-03/HINTS-2017-Consumer-Data-Brief-3.21.18.pdf. Accessed February 5, 2019

2 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010;363(06):501–504

3 Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012;157(07):461–470

4 Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and literacy independently predict patients' ability to identify out-of-range test results. J Med Internet Res 2014;16(08):e187

5 Pillemer F, Price RA, Paone S, et al. Direct release of test results to patients increases patient engagement and utilization of care. PLoS One 2016;11(06):e0154743

6 Fraccaro P, Vigo M, Balatsoukas P, et al. Presentation of laboratory test results in patient portals: influence of interface design on risk interpretation and visual search behaviour. BMC Med Inform Decis Mak 2018;18(01):11

7 Centers for Disease Control and Prevention (CDC). Colorectal cancer screening capacity in the United States. 2016 [cited June 14, 2018]. Available at: https://www.cdc.gov/cancer/dcpc/research/articles/crc_screening_model.htm. Accessed February 6, 2019

8 Sirovich BE, Welch HG. The frequency of Pap smear screening in the United States. J Gen Intern Med 2004;19(03):243–250

9 Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med 2018;33(04):533–538

10 Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behav Res Methods 2014;46(04):1023–1031

11 Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 2011;6(01):3–5

12 Peters E, Hibbard J, Slovic P, Dieckmann N. Numeracy skill and the communication, comprehension, and use of risk-benefit information. Health Aff (Millwood) 2007;26(03):741–748

13 Lalor JP, Wu H, Chen L, Mazor KM, Yu H. ComprehENotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation. J Med Internet Res 2018;20(04):e139

14 Short RG, Middleton D, Befera NT, Gondalia R, Tailor TD. Patient-centered radiology reporting: using online crowdsourcing to assess the effectiveness of a web-based interactive radiology report. J Am Coll Radiol 2017;14(11):1489–1497

15 Department of Health and Human Services. Innovative approaches to studying cancer communication in the new media environment (R21). 2018 [cited October 23, 2018]. Available at: https://grants.nih.gov/grants/guide/pa-files/par-16-248.html. Accessed February 6, 2019

16 Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgm Decis Mak 2010;5(05):411–419

17 Stewart N, Chandler J, Paolacci G. Crowdsourcing samples in cognitive science. Trends Cogn Sci 2017;21(10):736–748

18 Yank V, Agarwal S, Loftus P, Asch S, Rehkopf D. Crowdsourced health data: comparability to a US national survey, 2013–2015. Am J Public Health 2017;107(08):1283–1289

19 Dreyfuss E. A bot panic hits Amazon's Mechanical Turk. Wired. 2018 (August 17). Available at: https://www.wired.com/story/amazon-mechanical-turk-bot-panic/. Accessed February 6, 2019

20 Devine EG, Waters ME, Putnam M, et al. Concealment and fabrication by experienced research subjects. Clin Trials 2013;10(06):935–948