Original Paper

Authors: John Lalor¹ MSc, Beverly Woolf¹ Ph.D., Hong Yu¹,²,³,⁴ Ph.D. FACMI

¹College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA
²Department of Computer Science, University of Massachusetts, Lowell, MA, USA
³Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA
⁴Bedford Veterans Affairs Medical Center, Center for Healthcare Organization and Implementation Research, Bedford, MA, USA

Corresponding Author: Hong Yu, Ph.D. FACMI
Professor, Department of Computer Science
University of Massachusetts, Lowell
University of Massachusetts, Amherst (Adjunct)
Department of Medicine (Adjunct)
University of Massachusetts Medical School
and
Research Health Scientist,
Bedford VAMC
55 Lake Avenue North, AC7-059
Worcester, MA 01655
508-856-3474

Improving EHR Note Comprehension with NoteAid: A Randomized Trial of EHR Note Comprehension Interventions with Crowdsourced Workers

Abstract

Background: Patient portals are becoming more common, and with them the ability of patients to access their personal Electronic Health Records (EHRs). EHRs, in particular the free-text EHR notes, often contain medical jargon and terms that are difficult for laypersons to understand. There are many online resources for learning more about particular diseases or conditions, including systems that directly link to lay definitions or educational materials for medical concepts.

Objective: Our goal is to determine whether use of one such tool, NoteAid, leads to higher EHR note comprehension ability.

Methods: In this work we compare a passive, self-service educational resource (MedlinePlus) with an active resource (NoteAid) where definitions are provided to the user for medical concepts that the system identifies. We use Amazon Mechanical Turk (AMT) to recruit individuals to complete ComprehENotes, a new test of EHR note comprehension.

Results: Mean scores for individuals with access to NoteAid are significantly higher than the mean baseline scores, both for raw scores (p=0.008) and estimated ability (p=0.017).

Conclusions: In our experiments we show that the active intervention leads to significantly higher scores on the comprehension test as compared to a baseline group with no resources provided. In contrast, there is no significant difference between the group that was provided with the passive intervention and the baseline group. Finally, we analyze the demographics of the individuals who
participated in our AMT task and show differences between groups that align with current understanding of health literacy between populations.

**Keywords:** Health Literacy; Crowdsourcing; Amazon Mechanical Turk; NoteAid; MedlinePlus

**Introduction**

**Background and Significance**

In recent years many hospitals have adopted patient portals to make medical records available to patients. In particular, patient portals allow patients to access their Electronic Health Records (EHRs). Patient access to EHRs can lead to positive health outcomes and greater patient understanding of their conditions [1–3]. However, EHRs and the progress notes that are included often contain complex medical jargon that is difficult for patients to comprehend.

Self-service educational materials are widely available, especially online. There is a wealth of information related to medicine and healthcare on the internet, ranging from well-maintained ontologies with curated educational materials to online discussion communities of patients that suffer from the same disease. With this information, patients with certain symptoms can find information about their condition on the internet. But is the wealth of information useful? That is, does simply having access to health information lead to better understanding? In this work we test the usefulness of both passive and active interventions for assisting patients with understanding medical concepts. The passive system, MedlinePlus [4], is an online repository of information and definitions for clinical concepts, diseases, and other terms related to healthcare. MedlinePlus is a large repository of high-quality healthcare information, but the user must search for the information that he or she is looking for. MedlinePlus does not automatically surface information for users.

NoteAid [5,6] is a freely-available online system developed by our team that automatically identifies medical concepts and displays their definitions to users. NoteAid has previously been shown to improve patient understanding of notes as measured by self-reporting [5,6].

In this work, our goal is to determine if access to NoteAid or MedlinePlus is associated with higher levels of EHR note comprehension. Do these interventions of educational materials improve a patient's ability to comprehend his or her EHR note? In this work we use the Amazon Mechanical Turk (AMT) crowdsourcing platform to give AMT workers (Turkers) the ComprehENotes EHR note comprehension test [7], a set of questions designed to test EHR note comprehension. Certain Turkers were not given one of the external resources, while others were provided with either MedlinePlus or NoteAid. Our results show that using NoteAid leads to significantly higher scores on the EHR comprehension test compared to the baseline population that was given no external resource. However, we found no significant difference between the Turkers with no resource and the Turkers who used MedlinePlus. Turkers were also asked to take the short Test of Functional Health Literacy in Adults (S-TOFHLA) to assess functional health literacy. All of the Turkers scored *adequate health literacy*, the highest level for the S-TOFHLA.

In this work we show that NoteAid has a significant impact on EHR note comprehension as measured by a test specific to that task. In addition, simply giving a patient access to sites such as MedlinePlus does not lead to significant improvements in test scores over a baseline group that had no external resources available to them. Finally, we analyze the demographics of the Turkers who completed our tasks. A regression model to predict test scores showed differences between demographic groups that align with current knowledge regarding health literacy. For example, individuals that reported education of less than high school scored lower than average, while individuals that identified as white scored higher than average.
Related Work

Health literacy is an important issue for patients. Low health literacy is a widespread problem, with only 12% of adults estimated to be proficient in health literacy [8]. The Institute of Medicine defines health literacy as “the degree to which individuals have the capacity to make appropriate decisions regarding their health [9]. Patients with low health literacy are often not able to understand instructions for medications from their doctors and cannot navigate systems for making appointments, filling prescriptions, and fulfilling other health-related tasks [10,11]. In addition, having low health literacy has been linked to negative health outcomes in areas such as heart disease and fear of cancer progression [12,13].

It is important to be able to test a patient's health literacy to identify those patients with low health literacy. Doctors can then provide these patients with educational materials to improve their understanding of medical terms and concepts. Testing health literacy is especially important with the proliferation of online patient portals, where patients can access their EHRs and EHR notes directly. If a patient cannot understand the medical concepts in their EHRs they will not be able to take action to improve their condition.

There are a number of tests for health literacy, among them the Test of Functional Health Literacy in Adults (TOFHLA) and the Newest Vital Sign (NVS) [14–16]. TOFHLA and its shortened form (S-TOFHLA) test comprehension and numeracy by providing scenarios to patients and constructing fill-in-the-blank questions by removing key terms from the scenario passages. NVS is a short test where patients are required to answer questions related to a nutrition label, to test whether the patient can navigate the label. These tests work well as screening instruments to identify patients who may have low health literacy, but they are broad tests and do not specifically test EHR note comprehension.

While these and other tests are available, the only test that specifically targets a patient's ability to comprehend their EHR notes is the ComprehENotes test [7]. The ComprehENotes test questions were developed using key concepts extracted from de-identified EHR notes. Questions were written by physicians and medical researchers using Sentence Verification Technique (SVT) and validated using Item Response Theory (IRT) [17,18]. The test set is the first of its kind that specifically tests a patient's ability to comprehend the type of content that is included in EHR notes.

Methods

In this we recruited Turkers on the AMT platform and asked them to complete the ComprehENotes EHR note comprehension test. Turkers were split into three groups and allowed to use one external resource when completing the test (or no resource in the case of the baseline group). Test results were collected and analyzed using Item Response Theory to estimate EHR note comprehension ability for each of the individuals, and group results were analyzed to determine if either of the external resources had a significant effect on test scores. Figure 1 illustrates our methodology at a high level. Details for each of the steps are described below.
Figure 1. Flowchart describing our experiment. AMT workers were randomly assigned to one of three tasks on the platform. They completed the ComprehENotes test with the use of the provided external tool. All scores were then collected, and ability estimated were obtained using Item Response Theory.

**Data Collection**

To assess EHR note comprehension we used the ComprehENotes question set [7]. The dataset consists of 55 questions to measure EHR note comprehension. Questions were developed by groups of physicians and medical researchers from de-identified patient notes, then filtered down to a final test set using IRT. The questions were provided to Amazon Mechanical Turk workers (called Turkers), who provided responses. These responses were used to fit an IRT model that estimated the questions ability to test EHR note comprehension. Of the original question set, 55 were retained as a test of note comprehension.

The ComprehENotes test set is most informative for individuals with low health literacy. That is, the standard error (SE) of the ability estimation is lowest at low levels of ability (e.g., -2 to -0.5). In addition, most of the ComprehENotes questions have low difficulty parameters. The difficulty parameters range from -2.2 to 0.7. That is, the questions are of a difficulty that individuals with lower than average ability have a 50% chance of answering the questions correctly. For example, if a question has a difficulty parameter of -1.0 then an individual with estimated ability of -1.0 has a 50% chance of answering the question correctly. Ability estimates are normally
distributed, so an individual with estimated ability of -1.0 is one standard deviation below the average individual. Textbox 1 shows two examples questions taken from the ComprehENotes test. Individuals are shown a snippet of text from a de-identified EHR note and asked to select the answer that has the same meaning as the bolded portion of the text.

Textbox 1. Samples questions taken from the ComprehENotes test.

<table>
<thead>
<tr>
<th>Instructions: Please read the following questions, making a note of the bolded text, and then examine the provided answer choices. Please select the answer that best represents the bolded portion of the question text.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amitriptyline 25 mg p.o. at bedtime, Bactrim 160 mg p.o. b.i.d. on Friday, Saturday and Sunday, hydrocortisone cream, <strong>pegfilgrastim 6 mg subcutaneous one dose</strong>. He will continue to return for scheduled chemotherapy and will also be following up with the Hematology/Oncology Clinic.</td>
</tr>
<tr>
<td>a. Do a under skin injection of one dose of 6 mg pegfilgrastim.</td>
</tr>
<tr>
<td>b. Pegfilgrastim 6 mg epidermal one dose.</td>
</tr>
<tr>
<td>c. Pegfilgrastim may prevent neutropenia.</td>
</tr>
</tbody>
</table>

**The patient is in for her physical examination today.** Overall, she is doing very well. She is not on any blood pressure medications at the moment, she is doing fine. She had some issues in the past, but that has settled down. Her blood pressure is 110/78 today on no medications, pulse 68 and regular, respirations 12. |
| a. No physical examination was performed for the patient today. |
| b. The patient came to do check her health. |
| c. An eye exam is not a part of a regular physical examination. |

We set up 3 AMT tasks for Turkers to complete. Turkers were presented with the ComprehENotes question set, one question at a time, and were asked to provide the correct answer. For one task (Baseline), the Turkers were instructed to not use any external resources when answering the questions. For the first treatment task (Treatment-MLP), Turkers were given a link to MedlinePlus and were told that they could use the site as a reference when completing the task. Turkers were encouraged to use the MedlinePlus page search functionality to search for definitions to unknown terms or concepts that appeared in the task. For the second treatment task (Treatment-NA), the Turkers were given a link to the NoteAid system and were instructed to use NoteAid to complete the task. The tasks were restricted, so that individuals who completed 1 were not eligible to complete the other 2. For all groups we collected demographic information about the Turkers age, gender, ethnicity, level of education, and occupation. We also administered the S-TOFHLA test for each group to assess functional health literacy and to compare S-TOFHLA and ComprehENotes scores.

Since we are not able to monitor the Turkers as they complete our tasks, we cannot know for sure that the baseline group did not use any external resources as instructed. However, we can be confident that they did not have access to NoteAid. To access NoteAid, the Turkers would have to have known the URL link to access the system, even though we did not provide it to them. Alternatively, the Turkers would have had to search for NoteAid online without knowing the name of the specific system we are testing. Therefore, we are confident that even if the baseline group did use some external source during the task, they did not have access to NoteAid. The baseline Turkers may have found MedlinePlus if they searched online for medical concepts during the task. For example, a Google search of “COPD definition” returns a MedlinePlus link on the first page. However, unless the Turkers knew about MedlinePlus before beginning the task it is unlikely that they would use MedlinePlus as a reference during the task.
We included quality control checks for our AMT tasks to ensure a high-quality response from the Turkers. First, we restricted access to our tasks to Turkers with a prior approval rating above 95% to include only Turkers whose work has been judged as high-quality by other requesters. We also restricted the task to Turkers located in the United States as a proxy for a test of English proficiency. Within the actual task, we included three quality-check questions, which consisted of a very simple question with an obvious answer. If any Turkker answered one or more of the quality control checks incorrectly, their responses were removed from the later analyses.

**NoteAid**

The NoteAid system supplies lay definitions for medical concepts in EHR notes [5,6]. Users enter the text from their EHR notes into the NoteAid system, which outputs a version of the note with medical concepts defined. When the user hovers his or her mouse over a concept, a popup with the definition is shown. Figure 2 shows a high-level overview of the components in the NoteAid system, with example text that has been annotated. Users enter their EHR note text into NoteAid, and are provided with a reproduction of the text, with key medical concepts linked to their definitions.

![Diagram of NoteAid system](image)

**Figure 2.** Example showing NoteAid simplified text.

NoteAid consists of two components. The *concept identifier* component processes input text and maps terms to medical concepts. The concepts are mapped to entries in the Unified Medical Language System (UMLS) using MetaMap [19,20]. It then filters the list of returned concepts to include only concepts that match a subset of possible semantic types related to patient health (e.g., disease or syndrome, lab or test result, etc.). The *definition fetcher* component uses the filtered list of concepts to pull definitions from an external knowledge resource (e.g., Wikipedia or MedlinePlus).
Previous evaluation of NoteAid has shown that patient self-reported comprehension scores improve when using the system [5,6]. However, there has not yet been an evaluation of NoteAid on a test of comprehension, as opposed to self-reporting scores.

**IRT Analysis**

The ComprehENotes test set was developed using IRT [18]. The test set was built according to a single factor, 3-parameter logistic IRT model with a fixed guessing parameter. The test therefore measures a single latent trait, specifically the ability to comprehend EHR notes. Once the model has been fit, ability for a new test respondent is estimated by estimating $\theta$ according to the respondent’s answers to the test questions after the responses have been converted to a correct/incorrect binary format. For a single test question $i$, the probability that individual $j$ answers the question correctly is a function of the individual's ability ($\theta$). Figure 3 includes three equations: Equation 1 is used to calculate the probability that individual $j$ with an estimated ability of $\theta_j$ will answer question $i$ correctly, Equation 2 calculates the probability that individual $j$ with estimated ability $\theta_j$ will answer question $i$ incorrectly, and Equation 3 calculates the likelihood of individual $j$’s set of responses $U_j$ to all items in the test set, where $u_{ij}$ is 1 if individual $j$ answered item $I$ correctly, and 0 if they did not.

Figure 3. Equations for Item Response Theory 3-parameter logistic models.

\[
p_i(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}
\]

\[
q_i(\theta_j) = 1 - p_i(\theta_j)
\]

\[
p(U_j|\theta_j) = \prod_{i=1}^{I} p_i(\theta_j)^{u_{ij}} q_i(\theta_j)^{1-u_{ij}}
\]

$p_i$ and $q_i$ are functions of the known item parameters, and therefore we can estimate $\theta$ via maximum likelihood for each Turker. We also calculated raw test scores for each Turker (percent of questions answered correctly) for comparison.

**Results**

**Turker Demographics**

We first report the demographic information for the Turkers who completed our tasks. Table 1 shows the demographic information that we collected from the Turkers for the Baseline, Treatment-MLP, and Treatment-NA groups. Overall, most of the Turkers who completed our tasks are white, young, and have at least an associate degree. In addition, most of the Turkers do not work in the medical field. These demographics are not representative of a wider population, and do not fit demographics that are more commonly associated with low health literacy [9]. However, our goal here is to compare the results with respect to different interventions. In this case we do not need to test individuals with low health literacy, we instead want to see if scores improve when users are provided with certain external resources.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Baseline</th>
<th>MedlinePlus</th>
<th>NoteAid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>27</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>14</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>Demographic</td>
<td>Baseline</td>
<td>MedlinePlus</td>
<td>NoteAid</td>
<td>Total</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
<td>-------------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Age</td>
<td>22-34</td>
<td>23</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>35-44</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>45-54</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>55-64</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>65 and older</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Asian</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Black/African American</td>
<td>8</td>
<td>3</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Hispanic</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>White</td>
<td>26</td>
<td>24</td>
<td>21</td>
<td>71</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>High School Diploma</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Associates</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Bachelors</td>
<td>20</td>
<td>14</td>
<td>14</td>
<td>48</td>
</tr>
<tr>
<td>Masters or Higher</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Nurse</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Med Student</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Other profession in medicine</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Other profession</td>
<td>36</td>
<td>25</td>
<td>22</td>
<td>83</td>
</tr>
</tbody>
</table>

**Influence of Interventions**

Our analysis includes both the raw test scores as well as the estimated ability level using IRT. Since the test set consists of questions that were fit using IRT, we can also calculate the ability of these Turkers, and test whether the mean ability score was higher for Turkers that used NoteAid. Ability is a useful metric as it takes into consideration which questions you answer correctly, not just how many. IRT models question difficulty, so by considering whether easy or difficult answers were correct, IRT allows for a more informative score than percent correct. For each Turker we calculated their ability score (θ) using the IRT model fit as part of the ComprehENotes data set [7]. We use the mirt and ltm open-source R packages for estimating [21,22].
Figure 4 plots the raw scores for each AMT Turker for our test set. The center rectangles span the range from the first quartile to the third quartile of responses, and the bolded line inside each box represents the median score. Open circles indicate outlier scores. The upper horizontal line marks the maximum score for each group, and the lower horizontal line is 1.5 times the interquartile range below the first quartile. As the figure shows, visually there is a spread between the populations that did and did not have access to the interventions. Median raw scores for the baseline and MedlinePlus groups are similar, while median scores for the NoteAid group is higher. The spread of responses for the treatment groups is also smaller than the baseline group.

Figure 4. Box plot of raw scores for baseline and treatment Turker groups. The treatment groups were able to use MedlinePlus and NoteAid, respectively, when taking the ComprehENotes test.

Figure 5. Box plot of ability estimates for baseline and treatment Turker groups. The treatment groups MLP and NA were able to use MedlinePlus and NoteAid, respectively, when taking the ComprehENotes test.
Figure 5 shows the box plots of ability estimates. Again, the median values for the baseline and MedlinePlus groups are similar and the median ability estimates for the NoteAid group is higher. The lowest ability estimates for the baseline and MedlinePlus groups are much lower than for the NoteAid group (2 standard deviations below the mean as opposed to 1 standard deviation below). This shows that even for individuals that use NoteAid and still struggle, the low range of ability is higher than when NoteAid is not used.

To test whether each intervention caused a significant difference in scores, we compared each intervention to our baseline using Welsh’s two-sample t-test. Table 2 shows the mean raw scores and mean ability estimates for Turkers in each group. Mean scores are significantly higher than the baseline for Turkers that had access to NoteAid, both with regards to the raw scores (p=0.008) and estimated ability (p=0.017).

<table>
<thead>
<tr>
<th>Group</th>
<th>Raw Score</th>
<th>Ability Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.831</td>
<td>-0.065</td>
</tr>
<tr>
<td>MedlinePlus</td>
<td>0.849</td>
<td>0.138</td>
</tr>
<tr>
<td>NoteAid</td>
<td>0.923*</td>
<td>0.477*</td>
</tr>
</tbody>
</table>

Regression Analysis

We also wanted to determine if demographic factors had an impact on test scores. To that end we fit a linear regression model to predict raw scores using demographic information and group (e.g., baseline or treatment) as features. The results of the analysis showed that the intervention (none/MedlinePlus/NoteAid) was a significant feature in predicting raw score. Also, certain demographic groups were significant in determining score. Regarding ethnicity, individuals who self-reported as White had a significant positive coefficient. Regarding education, individuals that have less than a high school degree had a significant negative coefficient. These results are consistent with what is known about populations that are at risk for low health literacy. Individuals with lower education often have higher instances of low health literacy, as well as minorities. Our populations for this task, particularly with regards to minorities and less educated individuals, was very small. Future work on NoteAid in minorities populations would be worthwhile to confirm these effects.

Comparison with S-TOFHLA

All Turkers who completed our tasks were also given the S-TOFHLA test to complete. Scores on S-TOFHLA place test-takers into one of three categories: inadequate health literacy, marginal health literacy, and adequate health literacy. It is most useful as a screening tool to identify individuals.
with low or marginal health literacy. All Turkers in our tasks were scored to have adequate health literacy. In fact, all Turkers either scored perfect scores or only answered one question incorrectly, while the scores from the ComprehENotes test covered a wide range of ability estimates. The ComprehENotes can be used to assess EHR note comprehension at a more granular as opposed to a screening tool such as S-TOFHLA, where the primary concern is identification of individuals with low health literacy.

**ComprehENotes Analysis**

Finally, we wanted to see if the IRT model that was originally fit as part of the ComprehENotes data set was validated by the response patterns that we collected from the Turkers. To this end, we selected the two questions that the most Turkers answered correctly as well as the two questions that the fewest Turkers answered correctly.

These questions can be considered the easiest and hardest, respectively, from our task. The difficulty parameters for these items as modeled by IRT match the expectation of how difficult these items should be. The two hardest questions from our task (in terms of how many Turkers answered correctly) have difficulty parameters of 0.7 and -0.3, while the two easiest questions have difficulty parameters of -1.8 and -1.4. The difficulty parameter is associated with the level of ability at which an individual has a 50% chance of answering the question correctly. So, the low difficulty levels imply that someone of low ability has a 50% chance of answering the question correctly. Conversely, a higher difficulty parameter means that someone must be of a higher estimated ability level in order to have a 50% chance of answering correctly.

**Discussion**

**Principal Results**

In this work we have shown the importance of targeted, active intervention when trying to improve a person's ability to comprehend EHR notes. By giving Turkers access to NoteAid, scores on the ComprehENotes test are significantly improved over a baseline population that had no external resources. On the other hand, Turkers that had access to MedlinePlus but had to search themselves for the information that they wanted did not have a significant improvement in scores. NoteAid automatically identifies key medical concepts and provides definitions, as opposed to the scenario with MedlinePlus, where a user must decide what to search for. The user may not know that a certain concept is key for understanding a passage, or they may assume that they understand certain concepts that they don't. By letting the user decide what to search for, important terms may be missed, and overall comprehension may be affected.

**Limitations**

There are limitations to this work. First, by using AMT we are not able to monitor the Turkers who complete our task to ensure that only the external resources that we provide were used. This is particularly true in the baseline group, where our expectation is that no external resource was used. However, it is unlikely that the baseline users were able to access NoteAid without prior knowledge of the system, so we can be confident that they did not use it in our task. If the baseline users did use external resources, they most likely used a passive resource such as Google or even MedlinePlus.

In addition, the demographics of the Turkers who completed our task are not representative of the larger population, specifically among demographics associated with higher risks of low health literacy. In the case of this work that is not problematic, since our goal was to examine the effect of active and passive interventions on EHR note comprehension. The demographics of our three groups were similarly distributed, so the changes in scores can be linked to the intervention used.
Future work should look specifically at groups associated with low health literacy to determine if our results hold for those groups as well.

**Conclusions**

In this work we have shown that simply having access to resources designed to improve health literacy and medical concept understanding is not enough to provide benefit. The Turkers in our experiment who had access to MedlinePlus did not score significantly higher on the ComprehENotes test than those Turkers that were not provided with an external resource. On the other hand, having access to NoteAid, which actively pulls definition information and provides it to the user, led to significantly higher scores for Turkers. This result validates previously reported self-scored comprehension results showing that users had an easier time understanding their notes when they had access to NoteAid.

Knowing that users do not see benefits from simply having access to MedlinePlus is an important observation. When doctors are recommending next steps for patients who wish to improve their health literacy, it is not sufficient to point them to online resources. Targeted interventions are necessary to ensure that patients are able to learn about specific concepts and diseases that are relevant to them.

There are several directions for future work. Developing target curricula is necessary to ensure that patients can see benefits from online resources. They may not need a tool such as NoteAid (e.g., if they are not looking at their notes), but something more targeted than MedlinePlus is needed to ensure that patients are learning. Also, there should be further validation of the ComprehENotes test set with patients that are at risk for low health literacy. The Turkers in our task all scored either close to average or above average in our ability estimates, except for a few outliers. The test was designed to be most informative for individuals of lower ability, so this test should be replicated with such a population.

**Acknowledgements**

**Conflicts of Interest**

**Abbreviations**

**References**


