Will Gen Z users look for evidence to verify QA System-generated answers?

Soumya Gayen, Dina Demner-Fushman and **Deepak Gupta**
National Library of Medicine, NIH, HHS
{gayens,ddemner,guptadk}@nih.gov

Abstract

The remarkable results shown by medical question-answering systems lead to their adoption in real-life applications. The systems, however, may misinform the users, even when drawing on scientific evidence to ground the results. The quality of the answers may be verified by the users if they analyze the evidence provided by the systems. User interfaces play an important role in engaging the users. While studies of the user interfaces for biomedical literature search and clinical decision support are abundant, little is known about users' interactions with medical question answering systems and the impact of these systems on health-related decisions. In a study of several different user interface layouts, we found that only a small number of participants followed the links to verify automatically generated answers, independently of the interface design. The users who followed the links made better health-related decisions.

1 Introduction

The 2022 Health Information National Trends Survey highlighted the pervasive presence of health misinformation in social media and particular vulnerability of younger adults (18-34) to it (Chandrasekaran et al., 2024). Misinformation generated by Large Language Models (LLMs), referred to as hallucinations, is a known problem that instigated research in approaches that require LLMs provide references for each fact stated in the answer. A community-wide evaluation of the evidence provided by LLMs to support answers to medical questions shows that some of the provided references are irrelevant, do not support or even contradict the answer statements (Gupta et al., 2024). Having a question answering system to provide evidence is, therefore, not enough: it is also important to provide easy access to evidence and encourage its exploration through user interface design (Hullman and Gelman, 2021).

While research on interface design to support clinical decisions is substantial, it mostly addresses supporting clinical workflows and, based on many studies, recommends minimizing cognitive load by reducing the number of mouse clicks, among other approaches (Miller et al., 2018). Our objective, however is to find a layout that may encourage the users to drill down and analyze the evidence, i.e., increase the click-through without overwhelming the users. A study of strategies that ensure the users remain engaged with mobile phone health applications showed that the number of clicks increased due to content and graphics, among other factors (Moungui et al., 2024). Similarly to our goals, medical conversational agents are interested in keeping the users engaged. A recent review on artificial intelligence-based question-answering systems in health care, however, found that more is reported on the systems' effectiveness, and less is known about their use (Budler et al., 2023).

In this work, we explore several UI/UX design choices to determine if highlighting access to evidence leads to better use of evidence and, subsequently, better health-related decisions. Specifically, we studied if interleaving the links to evidence with answer sentences and highlighting the links with graphics, as well as making the images illustrating the answers more visible and clickable will lead to increased engagement of the users. In addition, after reviewing the answer and the evidence provided for a given health-related scenario, the users were asked to make a health-related decision or answer a health-related question on the topic of the scenario.

The results of this pilot study show an alarming tendency among the young and well-educated users with fair health literacy levels to blindly accept the displayed answer and subsequently make suboptimal health-related decisions. Only about a third of the study participants explored the evidence. These participants made better health-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 228–235
May 4, 2025

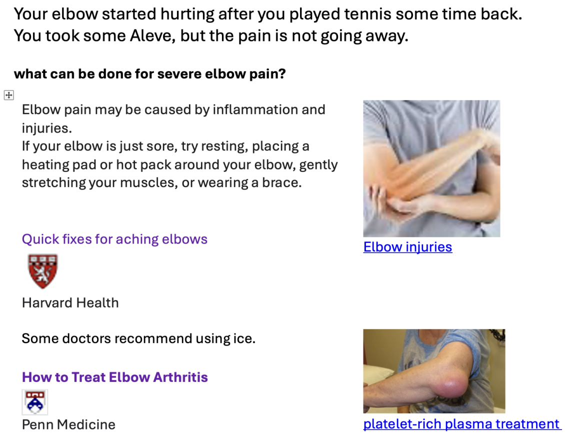

Figure 1: A user interface design with answer sentences interleaved with references, vertical figure bar, and links made more prominent using icons.

	Scenario							
User	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	A	B	H	C	G	D	F	E
2	B	C	A	D	H	E	G	F
3	C	D	B	E	A	F	H	G
4	D	E	C	F	B	G	A	H
5	E	F	D	G	C	H	B	A
6	F	G	E	H	D	A	C	B
7	G	H	F	A	E	B	D	C
8	H	A	G	B	F	C	E	D

Table 1: Different interface types used in 8x8 Latin square design. **Conditions** – *text*: blob (TB), sentence-by-sentence (TS); *pictures*: Vertical (PV), Horizontal (PH); *links*: Text (LT), icons (LI). **A**: TB, PV, LT; **B**: TS, PV, LT; **C**: TB, PH, LT; **D**: TS, PH, LT; **E**: TB, PV, LI; **F**: TS, PV, LI; **G**: TB, PH, LI; **H**: TS, PH, LI

related decisions.

2 Methods

To eliminate biases introduced by the order of presentation of the layouts and health scenarios, we chose the Latin Square design for our experiment (Richardson, 2018). We developed eight health scenarios containing a question, reference answers composed using reliable sources, and found relevant images linked to evidence using an image search engine. Using these scenarios, we studied two variants each of 1) answer layout & link placement, 2) image placement, and 3) augmenting the links with icons, making it 8 different types of interface from A to H. The answer was displayed as a paragraph followed by the references, or sentence by sentence interleaved with references as shown in Figure 1. The related pictures were shown in a horizontal or vertical image scroll bar.

We recruited eight students from a convenience sample of summer interns in the age range found most vulnerable in the 2022 Health Information National Trends Survey. Their educational background ranged from incoming college freshman to graduate level. A health literacy evaluation of the participants was performed to assess their medical data interpretation skills. This evaluation was performed in a classroom setting with limited time, to capture most accurate user health literacy information about the participants. We have used the test designed by Schwartz, Woloshin and Welch to establish the basic attributes, reliability and validity of a medical data interpretation test in a group of people with a wide range of quantitative abilities (Schwartz et al., 2005). In their experiment, the scores were normally distributed with a mean score of 61 and standard deviation of 17. Based on this mean score and the scores in our test, we divided the participants into 3 bins with score ranges 0 to 43, 44 to 78 and 78 to 100.

After completing the health literacy test, the students were given access to a web-based evaluation interface that displayed the eight questions according to the random Latin Square shown in Table 1. The questions were selected to reflect three levels of difficulty: factoid questions, questions about treatment effects, and information needed to support clinical decisions. The participants were instructed to read the scenario, and explore the answer and the presented evidence until they believed they could act on the information. In the next screen, they were presented with multiple choice answers / actions, from which they had to select one. For example, for the scenario shown in Figure 1, the choices where: a) *Give your elbow some rest, apply hot or cold, take more painkillers.* b) *Ask your doctor for advice.* c) *Ask your doctor for steroid injection.* d) *Ask your doctor about the experimental treatments such as acupuncture.* e) *Ask your doctor to refer you to see a surgeon.*

During the evaluation, all user actions were captured by the interface. Interactions, such as link clicks to patient-oriented reputable websites, data popup clicks (which displayed the original scientific publications corresponding to the patient-oriented materials accessible through the links), and related image scrolls were captured. Number of links clicked by the participants were recorded.

Time spent on every question by participants was also captured. After completing all eight scenarios, the participants completed a survey.

The survey asked which parts of the presented evidence informed the user's answers to the questions and decisions for immediate actions. It also asked if the answers were supported by the provided evidence and if the user felt a need to verify the answer before acting on it. Finally, the survey asked if the users would change any of the answers to the above questions if they knew the whole process was automated. After the study, the preferences for the page layout were discussed in the focus group with study participants.

2.1 Data Analysis

We assessed the responses to the selection of multiple choice answers/actions for a given scenario in two ways. In a strict evaluation, participants were awarded 1 point for each correct answer and 0 points for incorrect answers. Since the second choices for most questions are also reasonable, in a more lenient evaluation, the best answers received 2 points, while the second-best answers were assigned 1 point and the other answers received 0 points. We used Analysis of Variance (ANOVA) python package (Seabold and Perktold, 2010) for three factor design to analyze the effect of participants, questions, and interface types on use of evidence.

Source	SS	DF	F	Pr(>F)
Participants	1.11	7	1.18	0.34
Questions	7.86	7	8.34	0.001
Interface	1.36	7	1.44	0.21
Residual	5.66	42	NA	NA

(a) ANOVA results for strict evaluation of health-related decisions.

Source	SS	DF	F	Pr(>F)
Participants	2.67	7	1.28	0.29
Questions	7.94	7	3.78	0.002
Interface	2.19	7	1.04	0.42
Residual	12.63	42	NA	NA

(b) ANOVA results for lenient evaluation of health-related decisions.

Source	SS	DF	F	Pr(>F)
Participants	24103	7	2.99	0.01
Questions	13794	7	1.7	0.127
Time	6667	7	0.826	0.57
Residual	40682	42	NA	NA

(c) ANOVA table, results for time spent on every question by each participant.

SS:	Sum of squares
DF:	Degree of freedom
F:	F score
Pr(>F):	P value

Table 3: ANOVA results for strict and lenient evaluation of the use of evidence in health-related decisions.

3 Results and Discussion

User	Score	Group
1	67	2
2	33	1
3	44	2
4	67	2
5	78	3
6	72	2
7	56	2
8	78	3

Table 2: Health literacy scores.

Health literacy, defined as capacity to understand basic health information needed to make appropriate health decisions, was measured solely to mitigate the potential bias introduced by different health literacy levels. Our study participants were at least at the basic health literacy level, most of them were at the intermediate level, and two had high health literacy level as shown in Table 2. This finding agrees with the results of health literacy evaluation of college students that showed the university students seem to have good health literacy levels that would allow them to navigate the health care system (Ickes and Cottrell, 2010). The results of the literacy tests were not shared with the participants.

Table 3a presents the ANOVA results for the strict evaluation of the use of evidence in health-related decisions, while Table 3b shows the results for the lenient evaluation. In both evaluations, only the questions significantly affect the participants' decisions ($p = 0.001$ and $p = 0.002$).

On the aggregation of points scored by the participants, we find the Interface type C has achieved the highest scores (5 and 14) for both the methods of scoring. This suggests that participants could analyze and retain the data presented in this layout better. The focus group discussion confirmed that the participants preferred seeing the whole answer (rather than the individual facts interleaved with links to evidence), along with a horizontal image scroll bar, and the text only links to related research and clinical evidence. See Appendix D that shows the most and least popular interface designs.

Only 3 participants consistently clicked the links to patient-oriented evidence. Only 2 participants looked at scientific evidence (data pop-ups). Only one participant scrolled through the images on the screen. It shows that despite the preference for layout C, none of the layouts consistently engaged the users to drill through to the evidence. This suggests that the UI/UX we tested did not motivate the participants to check for evidence. Rather, the decision to seek supporting evidence was driven by their background knowledge, level of understanding, and confidence in the generated answers.

For the three users that engaged in interactions, we found a moderately positive correlation between the total number of user interactions and the score on health-related decisions. (See Appendix B). The participants who interacted more with the interface answered the follow-up questions better. Appendix C shows the amount of time spent by participants on the answer and evidence analysis before answering the follow-up question. ANOVA results in Table 3c show that variance in participants is statistically significant ($p = 0.01$), hence, the time spent on questions by every participant is not random, and a pattern is observed in user interactions. A moderately positive correlation in the amount of time spent on a question and score on the answers to follow-up questions and decisions was observed (see Table 4). It can be said that participants who spent more time reviewing the provided answers to the questions have answered the follow-up questions better.

The analysis of the exit survey results shows that all participants preferred information for patients, indicating a specialized patient-friendly system is needed. Only three participants did not trust the answer, they were the same participants that followed the links. This means that it's the application's responsibility to verify the correctness and accuracy of the user-facing information and ensure the information is absolutely trustworthy. This recommendation is reinforced by the fact that only one participant would make a distinction between the answers generated automatically and manually. The remaining seven participants indicated it doesn't matter how the answer is generated.

User	HL	SS	LS	Clicks	Time (ms)
1	2	5	13	12	2968
2	1	5	13	34	895
3	2	2	9	20	859
4	2	3	11	4	675
5	3	4	12	1	326
6	2	4	10	2	408
7	2	3	9	2	815
8	3	5	13	125	2409

Table 4: Users scores on the health-related decisions, their health literacy levels, and activity and time spent reviewing the answers to health scenario questions. HL - health literacy, SS - strict score, LS - lenient score.

4 Conclusion

Our study of the UI/UX designs for engaging users to verify the answers to their health-related questions shows that well-educated young adults with intermediate health literacy prefer seeing a full answer with unobtrusive links to supporting evidence and having illustrations below the answer. Studying the evidence provided to support the answers is associated with better scores on health-related decisions and medical topic understanding tasks. To confirm the association is significant and the results are generalizable, larger number of participants from more diverse population groups are needed. More studies are also needed to refine UI/UX design that engages the users and leads to optimal health-related decisions. Our results indicate that the majority of the users will not attempt to verify the answer reliability, which implies the onus of ensuring the correctness and accuracy of the answers is on the systems. The users who followed the links preferred reliable patient-oriented sources, which emphasizes the need for having such resources current, maintained, and curated by experts.

5 Future Works

In this pilot study, we experimented with only 8 users for the design choices of UI/UX to determine the appropriate way to highlight the evidence that may lead to better health-related decisions. In the future, we plan to extend the experiments with a more diverse and larger pool of users. We also plan to enrich the experimental setups with sophisticated tracking, such as eye gaze tracking (Wasfy et al., 2024), qualitative user feedback, and longitudinal studies (Kujala et al., 2011) to measure lasting behavior changes. We also plan to introduce more variables in designing choices by experimenting with different color schemes and font emphasis. To determine the usability of the UI/UX component, we plan to design a thorough questionnaire to assess the system Usability Scale (SUS) for better UI/UX designs of an effective QA system.

Limitations

This pilot study focused on a single age group. Although deemed vulnerable, the group is more technology savvy and better educated than many other population groups. To determine if the interaction patterns and health-related behavior displayed by this group is representative of the overall population, broader studies are needed. We hope that the study design and the evaluation interface code `https://github.com/soumyagayen/chqa-interface-evaluation` will help conducting more studies of the use of online medical question answering system and conversational agents.

Ethical Considerations

The patients' cases in this study were derived from the questions provided in the publicly available medical questions collections. The study participants volunteered and consented to participate in the study as part of their paid internship.

Data and code availability

All use cases, surveys and code are available at `https://github.com/soumyagayen/chqa-interface-evaluation`.

Acknowledgments

This research was supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. We thank the anonymous reviewers for the thorough reviews and excellent suggestions.

References

Leona Cilar Budler, Lucija Gosak, and Gregor Stiglic. 2023. Review of artificial intelligence-based question-answering systems in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1487.

Ranganathan Chandrasekaran, Muhammed Sadiq T, and Evangelos Moustakas. 2024. Racial and demographic disparities in susceptibility to health misinformation on social media: National survey-based analysis. *Journal of Medical Internet Research*, 26:e55086.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *Preprint*, arXiv:2411.18069.

Jessica Hullman and Andrew Gelman. 2021. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3):10–1162.

Melinda J Ickes and Randall Cottrell. 2010. Health literacy in college students. *Journal of American College Health*, 58(5):491–498.

Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. Ux curve: A method for evaluating long-term user experience. *Interacting with computers*, 23(5):473–483.

Kristen Miller, Danielle Mosby, Muge Capan, Rebecca Kowalski, Raj Ratwani, Yaman Noaiseh, Rachel Kraft, Sanford Schwartz, William S Weintraub, and Ryan Arnold. 2018. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. *Journal of the American Medical Informatics Association*, 25(5):585–592.

Henri Claude Moungui, Hugues Clotaire Nana-Djeunga, Che Frankline Anyiang, Mireia Cano, Jose Antonio Ruiz Postigo, and Carme Carrion. 2024. Dissemination strategies for mhealth apps: systematic review. *JMIR mHealth and uHealth*, 12:e50293.

John TE Richardson. 2018. The use of latin-square designs in educational and psychological research. *Educational Research Review*, 24:84–97.

Lisa M Schwartz, Steven Woloshin, and H Gilbert Welch. 2005. Can patients interpret health information? an assessment of the medical data interpretation test. *Med. Decis. Making*, 25(3):290–300.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Ahmed Wasfy, Nourhan Anber, and Ayman Atia. 2024. Eyeing the interface: Advancing ui/ux analytics through eye gaze technology. In *2024 Intelligent Methods, Systems, and Applications (IMSA)*, pages 538–543. IEEE.

Appendix

A Health-related decision and topic understanding evaluation results for each participant

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	1	0	1	1	0	1	0	1
2	1	1	1	1	0	0	0	1
3	1	0	0	1	0	0	0	0
4	1	0	1	1	0	0	0	0
5	1	0	1	1	0	1	0	0
6	1	1	1	0	0	1	0	0
7	0	0	1	1	0	0	0	1
8	1	0	1	1	0	1	0	1

Table 5: Strict evaluation results.

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	2	1	2	2	1	2	1	2
2	2	2	2	2	1	1	1	2
3	2	1	0	2	1	1	1	1
4	2	1	2	2	1	1	1	1
5	2	1	2	2	1	2	1	1
6	2	2	2	0	0	2	1	1
7	0	1	2	2	0	1	1	2
8	2	1	2	2	1	2	1	2

Table 6: Lenient evaluation.

B User behavior and interactions

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	4	1	2	1	1	0	0	0
2	0	5	2	0	0	2	1	0
3	19	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	2	0	0	0	0	0	0	0
7	0	0	0	0	0	1	0	0
8	23	9	7	6	10	7	5	0

Table 7: Number of links clicked by each participant on every question.

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	3	0	0	0	0	0	0	0
2	0	0	3	0	0	6	7	0
3	0	1	0	0	0	0	0	0
4	3	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0
8	4	6	0	0	5	6	0	0

Table 8: Number of popups (link to scientific evidence) opened by each participant on every question.

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	8	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	16	0	12	0	0	9	0	0

Table 9: Number of times the images have been scrolled by each participant on every question.

C Time spent on questions

	Scenario ID							
Participant	1	2	3	4	5	6	7	8
1	1040	87	1261	159	266	45	66	44
2	205	111	149	90	106	96	92	46
3	368	138	24	27	57	16	206	23
4	200	47	54	40	71	88	67	108
5	146	40	14	26	19	35	18	28
6	78	43	42	60	45	45	46	49
7	101	94	73	81	58	66	57	285
8	814	351	224	234	120	226	240	200

Table 10: Time spent by each participant on every question in seconds.

D Interface screen shots

Figure 2: Most popular interface Type C - (TB,PH,LT)

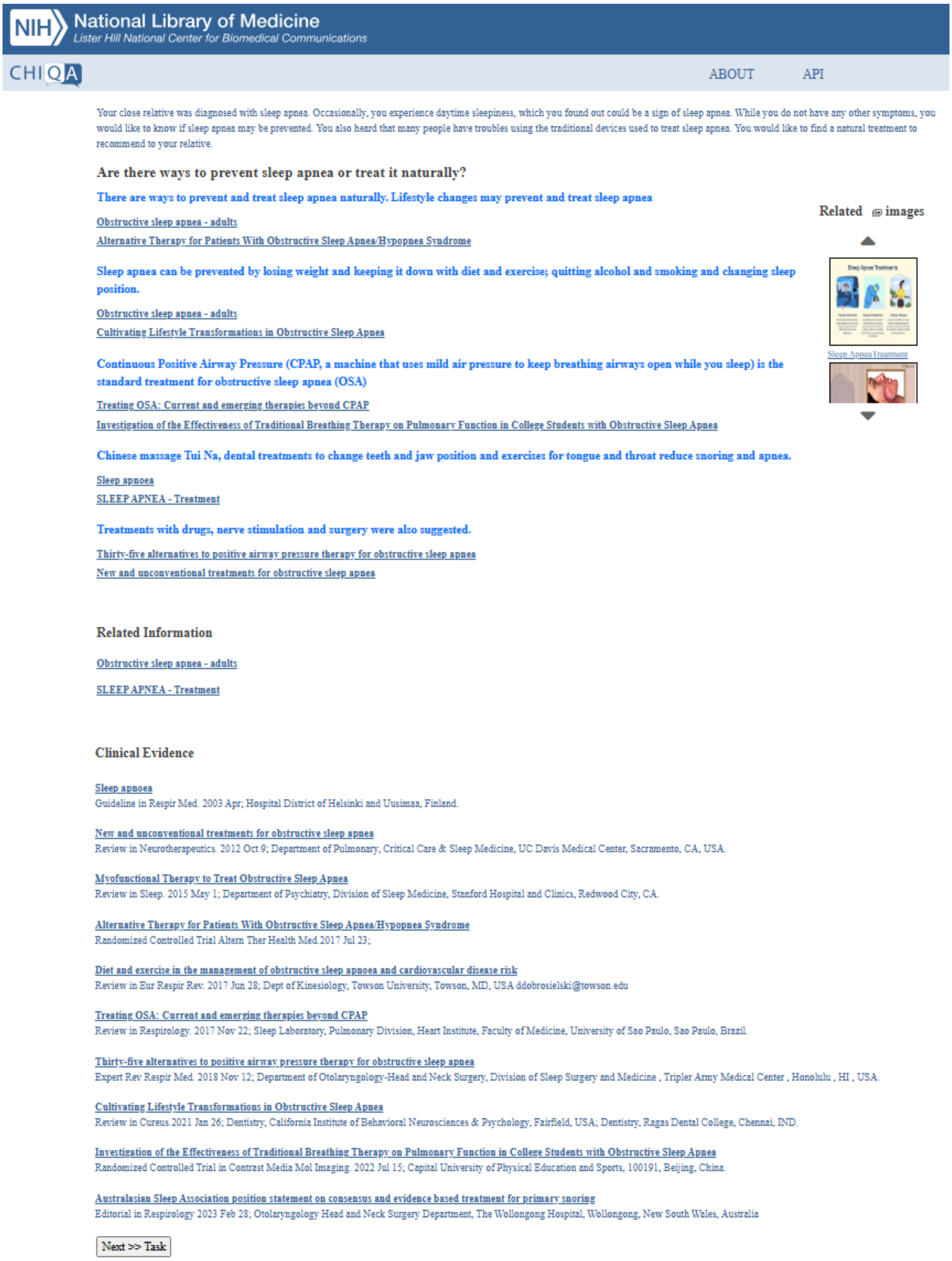

Figure 3: Least popular interface Type B - (TS,PV,LT)