A User-Centered Framework to Empower People with Parkinson's Disease

WASIFUR RAHMAN, University of Rochester, USA
ABDELRAHMAN ABDELKADER, University of Rochester, USA
SANGWU LEE, University of Rochester, USA
PHILLIP YANG, Center for Health + Technology, University of Rochester Medical Center, USA
MD SAIFUL ISLAM, University of Rochester, USA
TARIQ ADNAN, University of Rochester, USA
MASUM HASAN, University of Rochester, USA
ELLEN WAGNER, Center for Health + Technology, University of Rochester Medical Center, USA
SOOYONG PARK, University of Rochester, USA
E. RAY DORSEY, Center for Health + Technology, University of Rochester Medical Center, USA
CATHERINE SCHWARTZ, InMotion, USA
KAREN JAFFE, InMotion, USA
EHSAN HOQUE, University of Rochester, USA

We present a user-centric validation of a teleneurology platform, assessing its effectiveness in conveying screening information, facilitating user queries, and offering resources to enhance user empowerment. This validation process is implemented in the setting of Parkinson's disease (PD), in collaboration with a neurology department of a major medical center in the USA. Our intention is that with this platform, anyone globally with a webcam and microphone-equipped computer can carry out a series of speech, motor, and facial mimicry tasks. Our validation method demonstrates to users a mock PD risk assessment and provides access to relevant resources, including a chatbot driven by GPT, locations of local neurologists, and actionable and scientifically-backed PD prevention and management recommendations. We share findings from 91 participants (48 with PD, 43 without) aimed at evaluating the user experience and collecting feedback. Our framework was rated positively by 80.85% (standard deviation ± 8.92%) of the participants, and it achieved an above-average 70.42 (standard deviation ± 13.85) System-Usability-Scale (SUS) score. We also conducted a thematic analysis of open-ended feedback to further inform our future work. When given the option to ask any questions to the chatbot, participants typically asked for information about neurologists, screening results, and the community support group. We also provide a roadmap of how the knowledge

Authors' addresses: Wasifur Rahman, echowdh2@ur.rochester.edu, University of Rochester, Rochester, NY, USA; Abdelrahman Abdelkader, aabdelka@u.rochester.edu, University of Rochester, Rochester, NY, USA; Sangwu Lee, slee232@u.rochester.edu, University of Rochester, Rochester, NY, USA; Phillip Yang, Center for Health + Technology, University of Rochester Medical Center, Rochester, NY, USA, Phil.Yang@chet.rochester.edu; Md Saiful Islam, University of Rochester, Rochester, NY, USA, mislam6@ur.rochester.edu; Tariq Adnan, University of Rochester, Rochester, NY, USA, tadnan@ur.rochester.edu; Masum Hasan, University of Rochester, Rochester, NY, USA, m.hasan@rochester.edu; Ellen Wagner, Center for Health + Technology, University of Rochester Medical Center, Rochester, NY, USA, Ellen.Wagner@chet.rochester.edu; Sooyong Park, University of Rochester, Rochester, NY, USA, spark180@u.rochester.edu; E. Ray Dorsey, Center for Health + Technology, University of Rochester Medical Center, Rochester, NY, USA, Ray.Dorsey@chet.rochester.edu; Catherine Schwartz, InMotion, Beachwood, OH, USA, cschwartz@beinmotion.org; Karen Jaffe, InMotion, Beachwood, OH, USA, kjaffe11@gmail.com; Ehsan Hoque, University of Rochester, Rochester, NY, USA, mehoque@cs.rochester.edu.

2474-9567/2023/12-ART175
https://doi.org/10.1145/3631430

generated in this paper can be generalized to screening frameworks for other diseases through designing appropriate recording environments, appropriate tasks, and tailored user-interfaces.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **HCI design and evaluation methods**; **User studies**;

Additional Key Words and Phrases: Parkinson's Disease, End-to-end framework, Framework Evaluation

ACM Reference Format:
Wasifur Rahman, Abdelrahman Abdelkader, Sangwu Lee, Phillip Yang, Md Saiful Islam, Tariq Adnan, Masum Hasan, Ellen Wagner, Sooyong Park, E. Ray Dorsey, Catherine Schwartz, Karen Jaffe, and Ehsan Hoque. 2023. A User-Centered Framework to Empower People with Parkinson's Disease. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 175 (December 2023), 29 pages. https://doi.org/10.1145/3631430

1 INTRODUCTION

The present healthcare system is experiencing a digital transformation. It has significantly broadened its scope beyond the confines of hospital settings by facilitating remote screening and assessment of numerous diseases. The adoption of such digital screening technologies has been hastened in the wake of COVID-19, with the development of a variety of these tools for monitoring conditions such as eye diseases, cardiovascular illnesses, central sleep apnea, cognitive functions, hearing impairment, and depression, all from the comfort of one's home, to mention a few [38]. It's important that the widespread dissemination of these tools is matched with appropriate design considerations, aiming to empower users without causing any harm. Based on discussions with two neurologists, three People with Parkinson's (PwP), a director of a PD support group, and a User-Experience designer, we identified several design considerations including providing people with 1) clear instructions and guidance to do the tasks without any human intervention, 2) present assessment results in a coherent and intuitive manner, 3) opportunity to ask immediate clarifying and follow-up questions to prevent anxiety, and 4) a curated list of resources to empower them to take the appropriate follow-up actions. In this paper, we tackle these issues by developing a tool in partnership with the University of Rochester Medical Center, New York and subsequently assessing it through the In-Motion support group based in Ohio. The tool is designed for screening Parkinson's disease - the most rapidly increasing neurological disorder, for which remote evaluation holds significant promise.

Parkinson's disease (PD) – an incurable degenerative neurological disorder – is characterized by features such as slowed movement, decreased facial expression, and voice changes. Some of these symptoms are subtle and may only present themselves after a long period of time. Individuals diagnosed with PD at an early stage are more likely to comfortably manage their symptoms with existing medications and enjoy a good quality of life. The number of People with Parkinson's (PwP) doubled from 3 to 6 million between 1990 to 2015 and is projected to double again by 2040 [20]. Despite these dire statistics, screening and subsequent care for PD are complicated by issues of accessibility. In Beijing, China, almost half of PwP have never been diagnosed [21]. In the US, over 40% of diagnosed PwP with Medicare insurance do not or cannot visit specialists, even four years post-diagnosis [21]. Furthermore, movement disorder specialists typically practice in urban settings, which makes consultation particularly difficult for individuals living in rural and other underserved areas.

Prior work has demonstrated that it may be possible to develop a framework that allows anyone from anywhere to complete a set of neurological tasks – speech, facial expression, finger-tapping – and assess the likelihood of their demonstrating features of Parkinson's Disease (PD) [34]. However, it is important to use *user-centered* approach to validate whether the system can guide the participants to complete the set of neurological tasks seamlessly, present the results of the model to them in an intuitive and interpretable manner, and provide them an avenue to ask follow-up questions and actionable resources. In this paper, we focus on answering the following Research Questions (RQs) in the context of both the controls and the PwP:

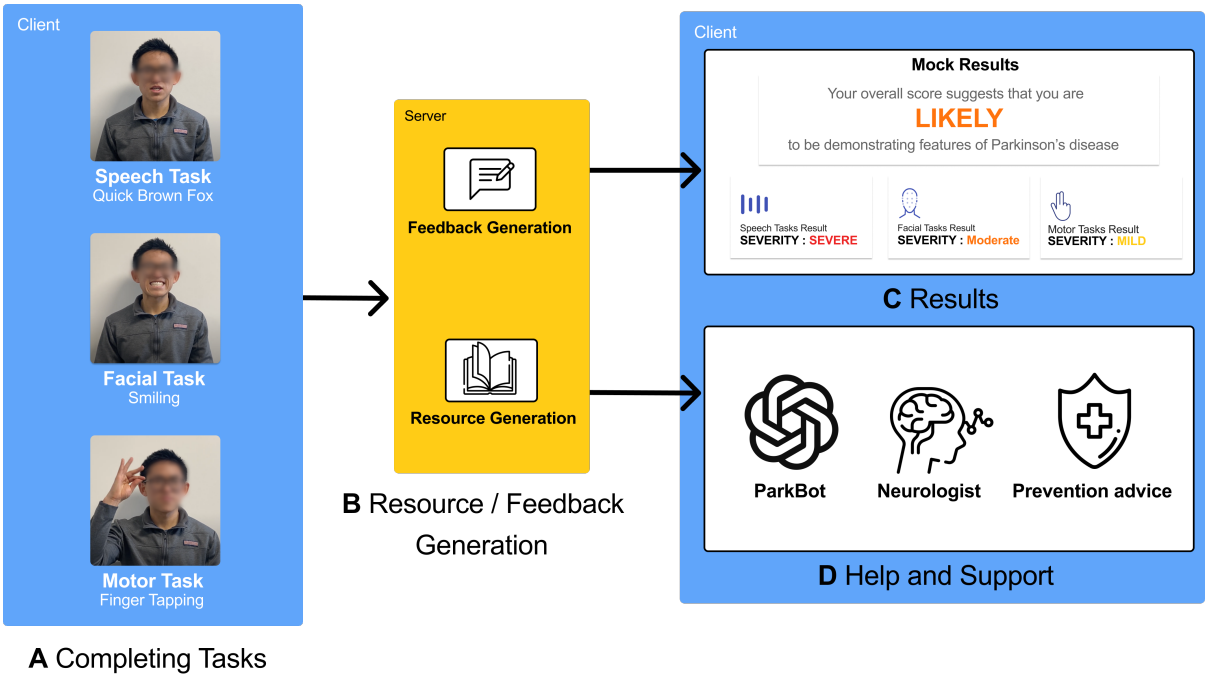

Fig. 1. End-to-end view of our framework for providing mock Parkinson's screening and resources. **A:** Users are guided to complete three sets of tasks to assess Parkinson's disease (PD) risk through speech, facial, and motor tasks. **B:** Recorded videos of them performing these tasks are passed to a server that processes them and generates mock predictions and feedback. **C:** This feedback, a screening assessment of PD risk and a breakdown of performance across the speech, facial expression, and motor domains, is displayed to users. **D:** The framework provides users with resources like a chatbot for real-time question answering about the screening outcome, highlighting local neurologists and support groups, and offering PD prevention and management advice. Parts of the figure are blurred due to HIPPA compliance.

- RQ1: How can we design a computer interface that allows participants to complete neurological tasks remotely without external supervision?
- RQ2: How can we visualize results from the output of AI models in a user-friendly way?
- RQ3: How can we allow participants to get answers to their follow-up questions immediately through the interface?
- RQ4: How can we provide actionable resources to the participants to help determine the next steps?

To answer these questions, we have built a framework consisting of three stages: 1) Tasks, 2) Results, and 3) Help & Support as outlined in Fig. 1. In the Tasks stage (Fig. 1.A), the participants can complete eight different neurological tasks – uttering "quick brown fox....", smiling, finger-tapping, etc. – divided into three domains: Speech, Facial, and Motor. Table 1 defines all eight neurological tasks. The chosen tasks form a subset of tasks from the previously validated PARK framework[34] and are consistent with the MDS-UPDRS guideline [26]. The framework guides a participant through completing each task by showing an instruction video and written instructions, an interface to record the task video, an opportunity to rewatch the instruction video, re-record the task, and appropriate corrective warnings (short video, microphone off, etc.). Once the videos are recorded, they are passed through a feedback system to generate some screening results (Fig. 1.B and Fig. 1.C). For this study, we generate separate pre-determined *mock* results for controls and PwP (Fig. 3). Once the participants see the

results, they receive four different items in Help & Support (Fig. 1.D): an AI-powered chatbot with GPT-3 backend; location of movement disorders specialist; actionable advice on diet, exercise, sleep, emotional well-being, and research participation; web-links to different organizations supporting PD research. Resources have been curated after several focus group discussions with neurologists and PwP to empower our participants to ask clarifying questions and get actionable advice so that they can take their next step with appropriate and scientifically proven information.

To validate our framework, we have designed a validation study consisting of four surveys (Fig. 5). After participants complete each of the Tasks, Results, and Help & Support stages, they take a survey designed to evaluate that particular segment. These three segments have a mix of ten multiple-choice and eight open-ended questions in total. In the end, we have a Final-survey consisting of ten multiple-choice questions to evaluate the usability of the entire framework by following the System-Usability-Scale (SUS) [9]. Table 2 outlines the questions in each stage. Across the first ten multiple choice questions (MCQs) in the first three survey stages, 80.85% of participants (standard deviation ±8.92%) rated it very favorably (e.g., either agree or strongly agree). We achieved median and mean SUS scores of 70 and 70.42(standard deviation ±13.85%) in the Final-survey stage, which is an above-average score (a SUS score above 68 is considered above average).

To analyze the open-ended questions, we organized the responses into multiple themes and provided a pathway to integrate those feedback into future iterations of the framework (6.1). Similarly, we analyzed the users' interactions with the chatbot using topic modeling and clustering techniques (6.2). We used the Latent Dirichlet Allocation (LDA) [8] for topic modeling to identify the main themes that emerged from the participants' chat history. In addition, we applied the K-means clustering algorithm [28] to identify patterns and trends in user messages. Our analysis shows that the participants were mainly interested in finding neurologists, communicating with the community support group that conducted the study, getting an interpretation of the screening results, etc.

To summarize, our main contributions are as follows.

(1) We developed a user-centered framework that can guide anyone with a computer with a webcam and microphone, from anywhere in the world, to perform a set of tasks as informed by the MDS-UPDRS (Section 3).
(2) We designed an interface that shares the PD screening results directly with users using an interpretable and patient-centered format (Section 3).
(3) We provided informative, well-reasoned resources that empower users to get answers through an AI-based chatbot, receive further evaluation, and learn about the actionable steps to slow the progression of PD (Section 3).
(4) We conducted a validation study with both PwP and control participants (Section 4) to evaluate the objective performance of the overall framework (Section 5). Moreover, We analyzed the open-ended survey responses (Section 6.1) and chatbot interactions (Section 6.2) to make our framework more user-centered.

2 RELATED WORK

Artificial Intelligence (AI) has transformed healthcare by achieving remarkable advancements in disease diagnosis and management. For instance, AI algorithms have demonstrated exceptional accuracy in analyzing mammograms and detecting early-stage tumors in cases of breast cancer [36]. Similarly, AI-powered systems can analyze retinal images to identify signs of diabetic retinopathy [5], enabling timely intervention to prevent vision loss. Additionally, AI shows promise in the early detection of Alzheimer's disease by analyzing cognitive patterns and brain imaging data [50], aiding in early intervention and treatment planning. However, the delivery of medical diagnostic results to patients poses a challenge, as it requires extensive training and experience to ensure that patients are empowered rather than anxious about their future. Therefore, thorough research is essential before

deploying AI models' predictions for any medical purpose. In the context of Parkinson's disease, although there has been a plethora of works involving diagnosing the disease with AI models [2, 3, 19, 29, 31, 34, 43, 46, 51], the question remains on how we can display the predicted results to the users, provide them with an avenue to ask clarifying questions and empower them with curated resources to seek further care. In this study, we provide a framework that investigates all these questions through the lenses of the users and suggests pathways for potential solutions.

2.1 Conveying Diagnosis/Screening Results to the Users

The effective presentation of diagnosis/screening results is crucial in healthcare as it directly influences patient understanding, decision-making, and follow-up actions. This is particularly critical in the context of AI-generated predictions. Researchers have emphasized the need for clear and understandable explanations to empower patients and minimize anxiety or distress. Strategies to mitigate user panic and anxiety include providing supportive information, offering resources for further understanding, and fostering open communication between healthcare professionals and patients [13].

Numerous studies have focused on developing user-friendly, intuitive, and informative approaches to display these results. For instance, Abukmail et al. [1] conducted a systematic review of research on visual presentations for communicating quantitative prognostic information to patients. The authors categorized eleven studies based on the type of presentation, which included bar graphs, pictographs, survival and mortality curves, tabular formats, and free text. The primary outcome assessed was the comprehension of the presented information. The study concluded that no single visual presentation consistently outperformed others in terms of patient comprehension, with results varying across studies and presentation types.

The study by Garcia-Retamero et al. [24] explored how visual aids can enhance patients' comprehension of medical risks and their self-assessment of understanding. The study involved 108 patients evaluating three medical tests, with and without visual aids alongside numerical data. The results revealed that without visual aids, many patients misunderstood test predictions and overestimated their comprehension. However, when visual aids were provided, patients across various levels of numeracy displayed high accuracy in making inferences and assessing their own understanding. These findings emphasize that well-designed visual aids can effectively communicate health risks and help patients gauge their level of understanding.

2.2 Evolution of Healthcare Bots

Healthcare chatbots have significantly improved in sophistication and usefulness due to advancements in natural language processing and improved computer hardware and software. These chatbots have evolved from simple communication interfaces to platforms capable of meaningful conversation-based interactions. For example, in 2015, Pharmabot [12] was developed to provide medication education for pediatric patients and their parents. This chatbot was designed to deliver accurate and easy-to-understand information about medications, thereby aiding in proper medication management. Another notable application was Mandy [41], a chatbot created in 2017 to automate the patient intake process in a primary care practice. By handling routine administrative tasks, Mandy freed up healthcare professionals' time to focus on more critical patient care activities. Similarly, in 2021, Nayak et al. [40] designed and developed a smart and efficient chatbot intended to serve as a daily health companion. This chatbot not only facilitates automated responses to user queries, saving time and simplifying daily life but it is also equipped to answer questions on a range of topics, including home remedies and the recent COVID-19 pandemic. The authors emphasized that the chatbot relies on well-vetted, reliable sources of information. It's worth mentioning that recent advancements in large language models have sparked significant research efforts aimed at harnessing the full potential of ChatGPT in the field of medical support chatbots [16, 27, 30, 45].

The remarkable capabilities of ChatGPT extend to supporting patients in managing their medications. It serves as a reliable companion, providing timely reminders, precise dosage instructions, and invaluable insights on potential side effects, drug interactions, and other vital considerations [37].

Thurzo et al. [48] undertook a comprehensive assessment of AI-based applications, including ChatGPT, in the dental domain. They deduced that ChatGPT could substantially enhance interactions between healthcare providers and patients, encompassing a range of tasks from analyzing patient messages to personalizing communication. However, they highlighted limitations in ChatGPT's ability to handle sophisticated tasks like understanding human anatomy. In a study conducted by Nov et al., ChatGPT's responses were compared to those of healthcare providers. Surprisingly, ChatGPT demonstrated a comparable rate of correct answers to the human professionals [42]. Jeblick et al. conducted a case study involving radiologists tasked with evaluating the quality of simplified radiology reports generated with ChatGPT. The results indicated that the reports were mostly accurate, complete, and not detrimental to patients. Nonetheless, the study did uncover some incorrect statements and missing medical information that could have potentially led to harmful conclusions. While highlighting the need for further improvements, the authors also emphasized the immense potential of ChatGPT in radiology [32].

Apart from medical applications, certain commercial symptom checker tools like Conversa [1] offer direct interaction with patients. These tools can ask relevant questions, comprehend symptoms, identify high-risk patients, and connect them to healthcare providers through in-person or virtual appointments. Wysa [2], a chatbot specializing in mental health support, offers evidence-based therapy exercises to help users manage symptoms of generalized anxiety and depression. Users have shown a strong emotional bond with the tool [7]. With 5 million users from 95 countries and a remarkable 91% user-satisfaction score, Wysa has proven to be highly impactful.

However, the use of chatbots like ChatGPT in medical writing and other applications comes with several limitations and ethical concerns [17]. These include issues of credibility, as the generated responses may lack proper validation or evidence, leading to potential misinformation. Plagiarism is another concern, as ChatGPT can bypass traditional detection methods, raising questions about the originality and authenticity of the content. Bias is a significant limitation, as ChatGPT learns from biased datasets, potentially perpetuating healthcare disparities and biases in decision-making. Lack of transparency is also a limitation, as it operates as a black box, making it difficult to understand how it arrives at its responses. Legal and medico-legal issues arise, such as copyright infringement and potential harm caused by inaccurate medical information. Privacy and data security are concerns, as ChatGPT relies on large amounts of data, raising questions about the protection of sensitive medical information. Additionally, ChatGPT may struggle with contextual understanding and may provide generic or inappropriate responses. It is crucial to address these limitations and ethical considerations to ensure patient safety, accuracy of information, and adherence to ethical standards in healthcare.

2.3 Principles and Guidelines for Responsible AI

Given the versatile usage and ever-growing momentum of Artificial Intelligence (AI) in all aspects of life, it has been stressed in recent times that any AI-based framework should be interpretable or explainable (XAI)[6]. In particular, Coalition for Health AI (CHAI) proposed a blueprint for trustworthy AI in healthcare [23]. Along with other studies[18, 49], it offers a guide for the development and deployment of AI systems in healthcare. The blueprint emphasizes transparency, fairness, robustness, user-centricity, and interpretability, among other principles, and underscores the necessity for these systems to meet rigorous ethical and professional standards. The study design should be transparent, with clear communication about the dataset curation, model design, and model performance. The individuals involved in the development, deployment, and maintenance of AI systems should be accountable for maintaining auditability, minimizing harm, reporting negative impacts, and

[1] https://conversahealth.com/
[2] https://www.wysa.com/

communicating design tradeoffs and opportunities for redress. The study should also be user-centric, with a focus on gathering user feedback to improve the framework. Additionally, the study should consider inclusiveness, ensuring that the framework is accessible to anyone, from anywhere. The design considerations in this paper are influenced by these new and emerging guidelines.

2.4 Our Contributions:

Based on the preceding discussion, we have pinpointed a gap in the existing literature regarding a comprehensive tool that can offer complete assistance to People with Parkinson's (PwP), encompassing task completion, result presentation, and resource provision. The present framework introduces the following innovative contributions in relation to both the state-of-the-art [34] and the aforementioned research gap:

(1) Adding three new facial expression tasks; improving the interface with new instruction videos, the capability to re-record each task multiple times, short-feedback messages to improve task recordings, etc.
(2) Providing 'mock' results to the users
(3) Providing access to a GPT-enabled chatbot to answer PD-related questions.
(4) Presenting a set of curated resources on diet, exercise, sleep, and mental health to improve the quality of life for PwP.
(5) Providing information about nearby neurologists, foundations, and communities supporting PwP.

3 DESCRIPTION OF OUR FRAMEWORK

We propose a user-centered framework consisting of three major components: (i) aiding the users in completing tasks with minimal cognitive load, (ii) presenting mock results to the users in an intuitive and interpretable manner, and (iii) offering a pathway to address any relevant questions the users may have and providing them access to actionable resources. We formed a multi-disciplinary team involving two neurologists, three PwP, a PD support group director, and a User-experience designer to identify these three components of the framework and their respective contents.

3.1 Framework Component: Completing Tasks

Table 1. Tasks in each of the domains. We divide the eight tasks into three domains: Speech, Facial, and Motor.

Domain	Tasks	Description of Tasks
Speech	Quick Brown Fox	Recite "The quick brown fox jumps over the lazy dog."
	Resting + Conversation	Rest for 10 seconds and describe the most recent book you read
Facial	Smile	Make a smiling face three times, alternated by a neutral expression.
	Disgust	Make a disgusted face three times, alternated by a neutral expression.
	Surprise	Make a surprised face three times, alternated by a neutral expression.
Motor	Finger Tapping (L)	Tap the thumb and index finger ten times for left hands
	Finger Tapping (R)	Tap the thumb and index finger ten times for right hands
	Extend Arms	Extend both arms fully for 10 seconds.

After participants log into our website (with registered email and password), they are able to read a short description of our entire framework (Fig 2.A). In our framework, we have integrated eight different tasks (organized into three domains). Table 1 provides a short description of each of the tasks along with the corresponding

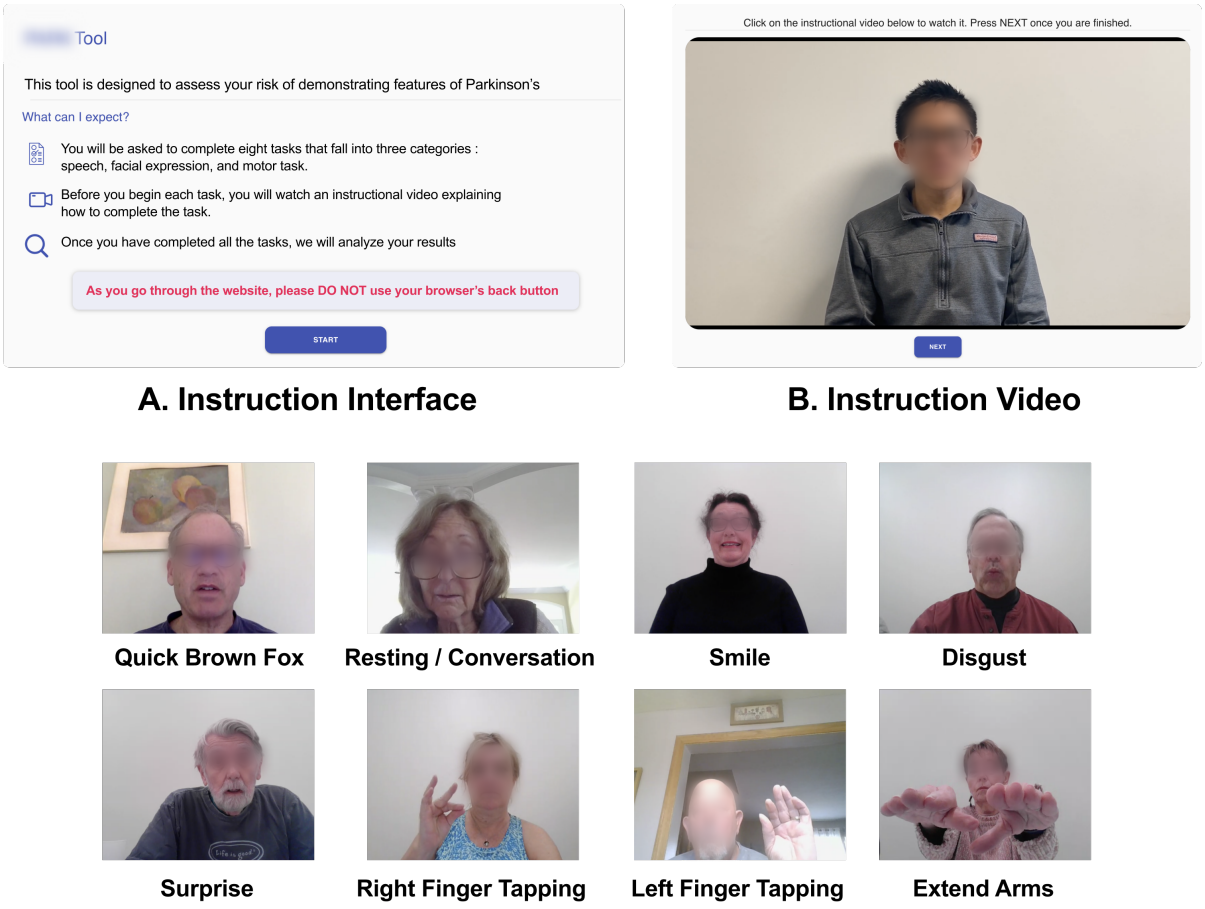

Fig. 2. Interface used for recording tasks and some recorded videos. **A:** participants can read a short description of the entire framework before starting to do them. **B:** For each of the tasks, the participants can watch a video explaining how to complete them. Then the participants can complete the tasks. Participants can re-watch the instruction videos and record a task multiple times until a sufficient video sample is recorded. Once participants are done recording a task, the system guides them to the next task until all are completed. **C:** We provide an example snapshot from the recordings of each of the eight tasks. Parts of the figure are blurred to adhere to double-blind review criteria.

domain name. While completing each of the tasks, the participants will be able to watch an instruction video (Fig. 2.B) and complete the tasks. Participants have the option to re-watch the instruction video multiple times and re-record themselves multiple times until they complete the task satisfactorily. We provide a screenshot of each of the eight tasks to show the diversity of the study participants (Fig. 2.C). To ensure quality data, our framework provides notifications (e.g., a recorded video is too short) and prompts the participants to re-record tasks if necessary.

3.2 Framework Component: Results

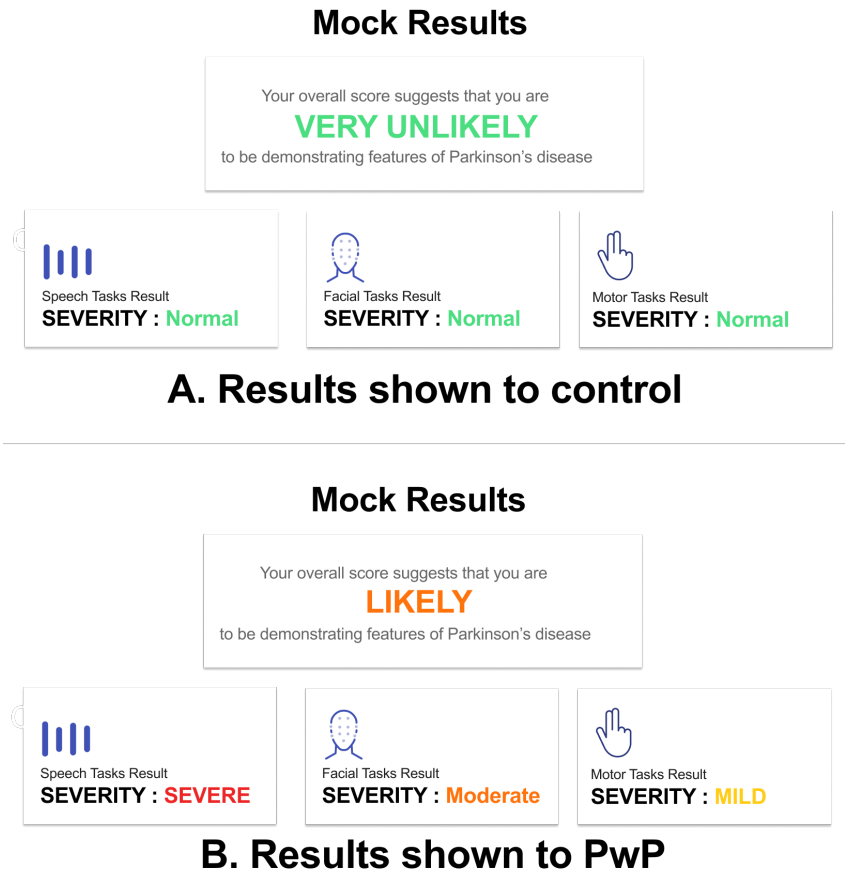

Fig. 3. The results we showed to Controls (**A:** Top) and PwP (**B:** Bottom).

.

Once the participants complete recording all the tasks, they are given a *Mock* result. Although we have integrated prediction models for speech [43] and finger-tapping tasks [2] that can give automated PD screening from the task recordings, we show *Mock* results to the participants of this study due to ethical considerations as explained in Section 4.2. Fig 3.A and Fig 3.B show the mock results for Controls and PwP respectively. For both PwP and Controls, we provide an overall result and then provide a breakdown of each of three domains: Speech, Facial, and Motor. In this study, each person with Parkinson's (PwP) receives an identical set of feedback, and similarly, each control group member receives a different set of identical feedback. If we were to present varying feedback to individuals within the same groups (PwP/control), we would need to assess how this variation affects their evaluation of the framework.

The language used to present the results was decided after several focused group sessions with multiple neurologists, PwP, and UX designers. The overall score is presented in a Likert scale (i.e., `very unlikely`, `unlikely`, `uncertain`, `likely`, `very likely`) since it is easy to understand. The domain-specific severity follows

MDS-UPDRS guidelines [25] for assessing PD severity, where the severity levels are `normal` (minimal or no symptom), `slight`, `mild`, `moderate`, and `severe` (most symptomatic). We chose these names for domain-specific severity for two reasons. First, since the MDS-UPDRS [25] is the current gold standard for diagnosing PD, these level names will be easily understood by the medical community. Second, we are training our models to predict the MDS-UPDRS severity level for each task video since the neurologists are well-trained to provide these ground labels. The overall score and the domain-specific severity tests were arranged in different fonts and colors to provide a visual aid for grasping the PD progression severity. Through focused-group sessions, we deduced that explanations of the overall score and the domain-specific severity scores visually can create clutters and confuse the participants. Instead, we provided an explanation of that score (adapted from the official MDS-UPDRS guideline [26]) to the interactive chatbot as input. If the participants wanted to get more detailed explanations of the scores, they could ask the chatbot for further explanations. In the future, our plan is to integrate models for all the tasks and generate the result automatically. By using the MDS-UPDRS severity levels, we can show our models' predictions without further transformations.

3.3 Framework Component: Help & Support

Once we show results to the participants, they are likely to have personalized queries about PD and will look for trusted resources to plan their next steps. Many of them might be unsure about what the scores mean, and what they imply. Looking through web search engines for these questions is not optimal because clinical answers found online are often too generic or sometimes misleading [14]. Upon receiving our screening results, we would like our participants to have the means to ask personalized questions about their screening results and condition and receive trusted answers and directions. With that goal, we designed a Large Language Model (LLM) based Chatbot for them to interact with and get answers to their personalized questions. The Chatbot is built on top of the GPT-3 TEXT-DAVINCI-003 model from OpenAI [10], which is primed to provide helpful information about Parkinson's disease and the participant's screening scores. The Chatbot can retain the memory of the previous 4 turns of conversation, which allows the user to refer to earlier dialogue. The Chatbot is instructed with the following goal,

(1) Explain the user's scores in Speech, Facial, and Motor tasks.
(2) Provide helpful guidelines related to Parkinson's disease, such as exercise, diet, sleep, etc.
(3) Provide localized directions to neurologists, support groups, and physical therapists based on the participant's geolocation.
(4) Making the participants aware that this is not an official diagnosis of PD, and such a diagnosis must come from an expert clinician.

However, Large Language Models are often prone to hallucinate and provide incorrect information on factual or time-sensitive information [4]. To prevent that, we prime the LLM to abstain from providing any information on the following list. Instead, the Chatbot will direct the user to a trusted website where they can find that information.

(1) Sensitive clinical suggestions.
(2) Recommendation on any medications.
(3) Information about specific neurologists, clinics, or information that are time-sensitive.
(4) Provide any information that can be harmful to the participants.
(5) Engage in off-topic conversation outside the scope of Parkinson's Disease.

The full prompt used to prime the chatbot is detailed in Appendix A. Fig 4 provides some screenshots of participants' interaction with the Chatbot. As Fig 4.B shows, the chatbot can respond to provocative questions

like whether the participant will die from PD. Fig 4.C and Fig 4.D shows our guardrails in action. Instead of suggesting medications or neurologists directly, we provide links from Parkinson.org or Google Maps.

Besides access to the chatbot, we also provide resources for improving the quality of life for PwP. Previous research shows that maintaining a healthy diet, regular exercise regimen, adequate good-quality sleep, and healthy emotional state can greatly improve the quality of life for PwP [22]. Therefore, we provide the participants with curated and validated resources on maintaining a healthy diet, exercise habits, sleep quality, and emotional state. We curated these resources from the book "Ending Parkinson's Disease: a Prescription for Action" [22], augmented them with references from well-established sources, and consulted with neurologists and UX experts. Fig 12 (in appendix) provides some examples of the provided resources: Fig 12.A, Fig 12.B, Fig 12.C provide example advises on the diet, exercise, participation on research. Fig 12.D provides clickable links to different foundations and communities supporting PwP.

4 VALIDATION STUDY

In this section, we a) describe our four-phased experimental design, b) provide the reasoning behind showing mock results instead of predictions from models, and c) narrate how we recruited and retained our participants.

4.1 Experimental Design

To validate our framework, we have designed an experiment consisting of the steps as outlined in Fig. 5. Our framework can be thought of in three parts: Tasks, Results, and Help & Support. After each of these steps, there is a survey to validate that particular part. Then, we have a final feedback stage to determine the usability of the entire framework. Table 2 lists all the questions asked in each of the survey stages. All the questions except those colored in blue (open-ended questions) are answered on a five-scale Likert scale: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree. The *Final-survey* portion is inspired by the System Usability Scale (SUS) [9] designed to measure the usability of the entire framework. A SUS score above 68 is considered above average. The open-ended questions (in blue-colored text) are designed to elicit suggestions to improve the framework in the future.

4.2 Ethical Considerations

In our study, we enrolled both PwP and healthy controls. It could be tempting to design a fully automated prediction engine and recruit participants to validate the prediction and the way it presents and explains the output of the classifier to the participants. It is possible for prediction models to make mistakes yielding an incorrect assessment. Incorrect predictions can cause severe anxiety and mental trauma to our participants. If PwP receives an assessment informing them of low disease risk, this can create a false sense of security and lead them to question prior diagnoses and/or become skeptical of current treatment regimens. On the other hand, if Controls receive an assessment informing them of high disease risk, this can create unnecessary stress, as well as create fiscal and logistical burdens. Moreover, neurologists are trained to share negative results with empathy to comfort the patients upon receiving potentially distressing news. Because our framework is not equipped to have participants discuss negative results with a trained professional, it may create a significant mental burden.

As a result, in this paper, we have separated the prediction engine and the screening information in the validation process. In the future, we plan to integrate the prediction engine and the screening information, while ensuring that patients will have the ability to talk with neurologists immediately, should they want to. This design minimizes any potential harm to the users and complies with the IRB guidelines.

Table 2. The survey questions asked to participants are listed below (best viewed in color). We conducted the survey in four phases: *Tasks-survey*, *Results-survey*, *Help & Support-survey*, and *Final-survey*, as outlined in Fig 5. Questions belonging to each phase are grouped together in the table. All but the questions written in blue are multiple choice questions (MCQ), with five possible answers: strongly disagree, disagree, neutral, agree, and strongly agree. The questions colored in blue are the only open-ended ones. Some words were redacted with an 'X' to adhere to double-blind review criteria.

ID	Question
	Tasks-survey
Q-1	The instructional videos helped me to complete the tasks
Q-2	The tasks were easy to complete.
Q-3	What can we improve about the previous (Tasks) section in the future?
Q-4	If you have additional feedback on the previous (Tasks) section, please share it below.
	Results-survey
Q-5	The description of the risk of demonstrating features of Parkinson's disease was easy to understand.
Q-6	The description of the risk of demonstrating features of Parkinson's disease was informative.
Q-7	The severity ratings provided for the speech, facial expression, and motor tasks were easy to understand.
Q-8	The severity ratings provided for the speech, facial expression, and motor tasks were informative.
Q-9	What can we improve about the previous (Results) section in the future?
Q-10	If you have additional feedback on the previous (Results) section, please share it below.
	Help & Support-survey
Q-11	It was easy to interact with 'X'-Bot.
Q-12	'X'-Bot adequately answered my questions.
Q-13	How can we improve 'X'-Bot in the future?
Q-14	If you have additional feedback on 'X'-Bot, please share it below.
Q-15	The resources provided were informative.
Q-16	The resources provided were comprehensive.
Q-17	Are there additional resources we should provide to users in the future?
Q-18	If you have additional feedback on the resources provided, please share it below.
	Final-survey
Q-19	I think that I would like to use the 'X' tool frequently.
Q-20	I found the 'X' tool unnecessarily complex.
Q-21	I thought the 'X' tool was easy to use.
Q-22	I think that I would need the support of a technical person to be able to use the 'X' tool.
Q-23	I found the various functions in the 'X' tool were well integrated.
Q-24	I thought there was too much inconsistency in the 'X' tool.
Q-25	I would imagine that most people would learn to use the 'X' tool very quickly.
Q-26	I found the 'X' tool very cumbersome to use.
Q-27	I felt very confident using the 'X' tool.
Q-28	I needed to learn a lot of things before I could get going with the 'X' tool.

4.3 Recruiting and Retaining Participants

To recruit both PwP and healthy controls, we used two neurology registries maintained by our partner institutions: a community support group based in the rural Midwest and a major neurological center in the

Table 3. Demographic information of the 91 participants who completed the entire study. Among the 43 Controls and 48 PwP, 27 (62.79%) and 25 (52.1%) are female respectively.

PD status	Number	Mean age (Std deviation)	Median age
Controls	43 (27 Female, 16 Male)	62.51(14.79)	66
PwP	48 (25 Female, 23 Male)	68.71 (8.22)	69

Northeast. This registry includes the contact information of individuals who expressed a prior interest in participating in neurological studies. We sent a mass email to relevant registries with a short description of our study. We scheduled an in-person appointment or provided a web link to the participants. The in-person appointments were scheduled primarily for the members of the community support group to coincide with their physical visit. The in-person participants used laptops provided by us to complete the study since they were unlikely to bring their own laptops during their visit to the support group. The remote participants used their own computers to complete the study. Among our 91 participants, 52 completed the study in-person and 39 completed it remotely. The in-person study coordinator was instructed to set up the device, provide a short introduction, and let the participants complete the study themselves. After logging into our study website, they entered their email, whether they were diagnosed with PD in the past or not, their gender, and their age. Afterward, the system led them to the *Tasks-survey* stage of our experiment to begin the study.

Our study included four phases. Participants could opt not to complete all the phases and leave in the middle. Fig. 6 provides an overview of how the number of PwP and control participants, both male and female, changed across the different phases of the study. The number of participants completing each of the stages was (in order from the first phase to the fourth phase): 109, 103, 91, 91. Thus, 83.49% of our participants who started the study finished it. Each participant had the option of receiving a $15 gift card for completing the study through the fourth phase. Demographic information of the 91 participants who completed the entire study is provided in Table 3.

All our participants reside in the US and are comfortable with communicating in English. 87.91% (80/91) of our participants are white; the others are Black, Asian and mixed races, etc. Although we validated that all of our participants could record the tasks properly and complete the surveys – thus indicating a reasonable technical proficiency – we don't have any subjective/objective metric of that proficiency. Similarly, to reduce the study completion time, we didn't collect information regarding PD severity or symptoms (cognitive/memory-related issues), or the number of years since diagnosis.

While completing the study, the participants faced some problems and reported them to the moderator. Three such reports stand out:

(1) The window of recording turns black *very occasionally* on the participants' end – making it difficult for them to adjust their bodies. However, upon manual inspection, we found that the videos were recorded properly.
(2) Time to upload the videos can be long due to network latency – prompting some participants to keep clicking next.
(3) Participants suggested completing some of the tasks optional in case someone fails to do them.

The moderator did not intervene on the spot to solve these problems and informed us about them later.

5 ANALYSIS OF MULTIPLE CHOICE QUESTION RESPONSES

Across the four phases of our survey, 20 multiple-choice questions (MCQs) were asked in total; we asked 2, 4, 4, and 10 MCQs during the *Task-survey*, *Results-survey*, *Help & Support survey*, and *Final-survey* phases respectively. The participants responded to each MCQ with one of five answers: Strongly Disagree, Disagree, Neutral, Agree,

and Strongly Agree. For all the MCQ questions in the **first three** phases, higher percentages of Strongly Agree or Agree responses indicate that the participants responded to our framework positively.

To evaluate our framework, we designed three Evaluation Questions (EQs). The EQ1 and EQ2 focus on the first three phases since the fourth phase – **Final Survey** – is designed to evaluate the overall usability of the framework. The EQs are:

- EQ1: How did all the participants rate the first three phases of the framework?
- EQ2: Did both the PwP and controls rate the first three phases of the framework highly?
- EQ3: How suitable is the framework for all the participants?

If the participants provide high ratings to these three EQs, we will be able to show that we have answered the Research Questions (RQs) introduced in Sec 1 successfully. Across all three EQs, we show that most of our participants viewed our framework as user-friendly, informative, and useful.

5.1 Response to EQ1

To answer EQ1, we constructed a bar plot of the percentage of participants who responded with each of the five answer choices across each MCQ in the first three phases. We present the findings in Fig.7 and find that the median and mean of participants who responded to all our questions with either Strongly Agree or Agree to be 81.38% and 80.85% (std dev ± 8.92) respectively. Therefore, on average, 80% of the participants judged our framework favorably.

5.2 Response to EQ2

To answer EQ2, we put the responses from PwP and controls into two separate groups. Then, for each MCQ across the first three survey phases, we calculated the percentage of people in that group (PwP or Control) who answered either Strongly Agree or Agree. The results of that analysis are presented in Fig 8. Excluding Q-5 (p=0.007) and Q-7 (p=0.049), there isn't a statistically significant difference ($p < 0.05$) in the proportion of (Strongly) agree responses from PwP and the Control group (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$)

We found the median and mean of PwP participants who responded to all our questions with either Strongly Agree or Agree to be 75% and 77.5% (std dev ± 10.24) respectively. Similarly, we found the median and mean of control participants who responded to all our questions with either Strongly Agree or Agree to be 84.88% and 83.95% (std dev ± 8.66). Therefore, more than 75% of the participants in *both the PwP and control groups* judged our framework favorably.

5.3 Response to EQ3

To answer EQ3, we limited our analysis to the 10 MCQ questions (Q-19 to Q-28 in Tab 2) in the Final-survey phase. These questions are adapted from the System Usability Scale (SUS) [9] which is an industry standard for measuring system usability that can be administered very easily, works reliably for small sample sizes, and can easily differentiate between usable and unusable systems [3]. As we show in Table 4, all the SUS questions are answered on a 5-point Likert scale. Then, we transform those Likert scores into a numerical score, convert separately for the odd and even numbered questions since they have opposite polarity, sum the converted scores, and multiply the total score by 2.5 to get a final SUS score in the [0-100] range. In the end, we achieve median and mean SUS scores of 70 and 70.42(std dev ± 13.85), which indicates that our system has above-average user-friendliness.

[3]https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Table 4. Response options in SUS questions, their scores, conversion of scores for both odd and even-numbered questions. Those converted scores for all questions are summed up and multiplied by 2.5 to scale the total SUS score to the 0-100 range. A SUS score above 68 portrays an above-average user-friendly system.

Response option	Score	Odd question score (Score -1)	Even question score (5 -Score)
Strongly Disagree	1	0	4
Disagree	2	1	3
Neutral	3	2	2
Agree	4	3	1
Strongly Agree	5	4	0

We conducted additional analysis on the SUS scores obtained from PwP and the Control group. The median SUS scores for PwP and Controls are 71.25 and 70.0, respectively, showing no statistically significant ($p = 0.89$) difference (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$).

6 ANALYSIS OF OPEN-ENDED PARTICIPANT FEEDBACK

In this section, we conduct an analysis of both open-ended surveys and chatbot interaction history. The goal of the analysis is to get insights into the limitations of our framework and avenues to improve it.

6.1 Open-ended Survey Questions

In this section, we analyze participants' responses to open-ended questions in Tab. 2: Q-3, Q-4, Q-9, Q-10, Q-13, Q-14, Q-17, and Q-18. We group those responses into several themes using thematic analysis.

6.1.1 Tasks-survey.

(1) *Refining instructions*: A few participants provided suggestions for increasing instruction clarity. One PwP said, *"It would have been helpful to know if I should do it with or without glasses."*. Another PwP suggested improving the instruction videos: *"The video of the 'disgust' face seemed more like anger rather than disgust. I wasn't entirely sure how long to hold out my arms, and the instructional video could be made more effective."* Moving forward, we plan to implement these suggestions to make our instructions more intuitive and our framework more user-friendly.

(2) *Enhancing camera usage*: A control suggested providing a brief period for users to adjust the camera before recording begins: *"Allow for a second or two of screen time to adjust the camera before recording to make sure one is in the frame and on the last one, that your hands are able to extend."*. Based on the feedback received, we will implement a brief adjustment period in our framework before recording starts to improve the overall recording experience.

(3) *Timing with Medication Cycle*: Some participants expressed concerns about the timing of performing tasks in relation to their medication cycle. One participant commented, *"At what point in your med cycle should you be?"* This feedback highlights the importance of considering the participants' medication schedules. It might be beneficial to provide guidance to users about the best time to perform the tasks relative to their medication intake to ensure their safety and accurate and consistent data and results.

6.1.2 Results-survey.

(1) *Results Explanation*: A participant (from PwP) underscored the necessity of context in understanding results, suggesting the addition of *"explanations of what observations lead to the results."* Another participant (from PwP) noted, *"More clarity on the ratings would be beneficial, as interpreting feedback can be challenging when unsure about the task execution."* Reflecting on this feedback, we plan to provide clear explanations

of our screening tool's processes when we deploy real models, detailing how it extracts key features from the recorded videos and subsequently, how these extracted features contribute to the resulting assessment. Including such explanations not only adds transparency to the scoring system but also bolsters the acceptability of the system by providing clinically relatable interpretations.

(2) *Scoring System*: We received a suggestion from a PwP participant who believed a numerical scoring system might provide clearer and more objective feedback, enhancing understanding of the current rating system. We chose the current feedback by incorporating language from MDS-UPDRS feedback. But we will brainstorm further whether we can also incorporate a hybrid scoring system by adding both categorical (severe/moderate/mild/slight/normal) and numerical scoring.

(3) *User-friendly Language*. Finally, a participant (from PwP) reminded us of the importance of language choice in user experience, mentioning that the word 'severe' could potentially cause discomfort or worry. This suggestion will help us to refine our language to be more approachable and less alarming to users.

(4) *Objective metrics report for neurologists*: One participant raised a valuable question,*"Is there something we can provide to PwP to take to their neurologist?"* This feedback highlights an opportunity to enhance the tool's utility by offering a feature that generates a comprehensive report or summary of the assessment results that could be reviewed by a neurologist for more insight.

6.1.3 Help & Support-survey.

(1) *Personalized Resources*. A number of participants highlighted the need for more tailored resources. One participant (from PwP) stated: *"The resources are highly beneficial for those with a confirmed diagnosis of PD. However, for screening purposes, it may be more helpful to have resources that are not exclusively disease-centric to avoid confusion or distress."*. In the future, we aim to further personalize our provided resources by offering a broader range of materials that are not exclusively centered on diagnosed PD – thus reducing potential confusion or distress during screening.

(2) *ChatBot Updates*. A participant (from the control group) emphasized the importance of the chatbot being current with new information and research. The participant positively noted, *"It is essential to keep the chatbot updated, which was reflected when it successfully answered a question about a recent biomarker research."* We are pleased to share that our chatbot has been designed with a focus on staying current with the latest information and research. We will continue to monitor its performance to maintain its relevancy and usefulness.

(3) *More Resources*: Multiple participants suggested adding more organizations to the resources section such as InMotion and the Parkinson's Buddy Network. Based on this feedback we plan to add more links to resources, books, and webinars about Parkinson's disease.

6.2 ChatBot Interaction Analysis

In this subsection, we provide a three-pronged analysis using visual inspection, topic modeling, and clustering analysis to understand interactions between the study participants and the chatbot. Our study is based on the chat history of our 60 participating individuals, who collectively posted a total of 116 messages to our chatbot. On average each individual roughly asked two queries. The primary goal of this analysis is to identify the key topics and concerns raised by users, which will enable us to make data-driven decisions for future improvements to the chatbot. Our analysis shows that the participants were mainly interested in finding neurologists, communicating with the support group that conducted the study, getting an interpretation of the screening results, etc.

To identify the thematic structure in the text data, we applied two different text analysis methods.

(1) **Latent Dirichlet Allocation (LDA) algorithm [8] for topic modeling**: Latent Dirichlet Allocation (LDA) is a type of probabilistic model, which assumes that every document (in our case, every chat message) is a mixture of a certain number of topics and that each word in the message is attributable to one of the document's topics. In our analysis, we used LDA to identify common themes within the chat history. The first step was text preprocessing, which involved cleaning the text, tokenizing, removing stop words, and performing stemming and lemmatization. After preprocessing, we converted the text data into a numerical form using a document-term matrix. We then applied the LDA algorithm to this matrix to extract topics. Determining the optimal number of topics was a crucial step; we used several methods including the elbow method, coherence scores, and manual inspection to select the best number. The LDA model was trained using a Term frequency-inverse document frequency (TF-IDF) [44] representation, and we interpreted the resulting topics based on the generated keywords. This comprehensive process allowed us to understand the key topics and concerns raised by the study participants during their interactions with the chatbot.
(2) **K-Means Clustering algorithm [28]** We also implemented K-Means clustering to categorize our chat data into distinct groups. After converting our text data into a numerical form using Tf-IDF vectorization, we used the elbow method [15] to determine the optimal number of clusters, 'K'. This method involves plotting the explained variation as a function of the number of clusters and picking the "elbow" of the curve as the appropriate number of clusters. With this optimal 'K', we ran the K-Means algorithm, which iteratively assigns each data point to one of the 'K' clusters based on distance from the cluster centroids. The resulting clusters, along with LDA-derived topics, provided us with a clear understanding of the themes in the chat history.

From both of these analyses, some themes and clusters arrive:

- Primarily focused on Parkinson's Disease and related resources, this topic covers inquiries about associated symptoms of PD, suitable medications and their prescribed dosages, and the current state of PD research with its findings and evidence. It also includes the search for neurologists, caretakers, and other supportive resources, which together constitute 31.6% of the total messages.
- Pertaining to concerns about participants' test results, which constitute 22% of messages. This includes questions regarding the detailed interpretation of results, understanding the severity level of each task, and so on.
- Comprised basic inquiries about *the study conducting community support group*, making up 11.7% of messages. This covered questions about its nearest locations, its available support services and the process for scheduling sessions, etc.

For visual inspection, we generated a word cloud for visualizing the most frequently mentioned words and phrases in the chat history with the chatbot (Fig 9). As shown in Fig 9, the most prominent terms in the word cloud correspond with the findings from our topic modeling and clustering analyses, with 'neurologist' being the most frequently mentioned word.

7 LIMITATIONS AND VALIDITY OF THE STUDY

7.1 Replication Limitation of Chatbot

Generative models such as GPT are renowned for their capability to generate diverse responses to a given question. In our study, participants had the liberty to ask any PD-related questions. Since we did not impose limitations on the range of permissible questions, we didn't examine the potential consequences of this lack of reproducibility. Nevertheless, we acknowledge the possibility that the chatbot's responses could be inconsistent and occasionally incorrect. To tackle this issue, we have introduced a general cautionary note for all participants:

"Please be aware that responses from X-Bot have not undergone formal verification and might therefore contain inaccuracies."

7.2 Technical Limitation of the Chatbot

During the development of our chatbot, we encountered the challenge of striking a balance between expressiveness, correctness, and safety. Generative models have the tendency to confidently produce incorrect responses [4], prompting us to prioritize correctness and safety over expressiveness (explained further in Section 8.3). This intentional decision has led the chatbot to provide responses that may seem generic, limiting its overall usefulness. However, looking ahead, we have devised a plan to enhance its expressiveness through two key steps.

Firstly, we aim to fine-tune the underlying GPT-3 model using verified PD resources from reliable sources like relevant books, neurological directories, and hospital websites. This will enable the chatbot to incorporate accurate and up-to-date information into its responses. Secondly, we intend to integrate another model into the chatbot's architecture. This additional model will leverage an external knowledge graph [39] to verify the correctness of the chatbot's responses. By cross-referencing with trusted sources, the chatbot can ensure that the information it provides is reliable and accurate. By implementing these measures, we aspire to create a more refined and capable chatbot that not only maintains its safety and correctness but also gains the ability to deliver more expressive and valuable responses to users.

7.3 Impact of Participant-Moderator Interaction

Our moderator worked primarily towards coordinating with the in-person participants, setting up the instruments, and letting the participants complete the study themselves. In order to quantify whether the moderator's presence had any impact on the user experience, we plot the proportion of (Strongly) agree responses from the In-person and Remote participants for each MCQ question in Fig 10. Excluding Q-15 (p=0.025) and Q-16 (p=0.023), there isn't a statistically significant difference ($p < 0.05$) in the proportion of (Strongly) agree responses from both groups (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$).

Furthermore, the median SUS score of both In-person and Remote groups is 70.0, showing no statistically significant ($p = 1.0$) difference (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$). Therefore, both the In-person and Remote groups enjoy statistically similar user experiences.

7.4 Interpreting the Domain-specific Severity Score

In our framework, we provided domain-specific severity scores `normal` (minimal or no symptom), `slight`, `mild`, `moderate`, and `severe` (most symptomatic) to the participants to adhere to the MDS-UPDRS criteria and our machine-learning paradigm. An alternative way to show the feedback is to through a percentage. However, after consulting with several PwP and neurologists, we found the percentage-based feedback to be quite difficult to interpret for the PwP. To validate our approach, we designed two research questions – Q-7 and Q-8 – in our user study:

- Q-7: The severity ratings provided for the speech, facial expression, and motor tasks categories were easy to understand.
- Q-8: The severity ratings provided for the speech, facial expression, and motor tasks categories were informative.

In response to Q-7, 75% of PwP and 90.69% of Controls responded (Strongly) agree. Similarly, while responding to Q-8, 75% of PwP and 83.72% of Controls responded (Strongly) agree. These scores suggest that our participants

found the severity ratings to be informative and easily understandable. Nevertheless, we concur that a more comprehensive study should be designed to compare various methods of providing feedback and to assess the trade-offs involved in adopting one approach over another.

7.5 Framework Usability and Applicability

The core components of our framework – completing tasks, getting feedback, and resources – can be explored with just a computer mouse/touchpad without using the keyboard (interacting with the chatbot is optional). Since webcams and microphones are built into the laptops, they are the most ideal medium for interacting with our framework. Using our framework with a desktop requires setting up an external webcam and microphone.

Similarly, the framework can be more beneficial with early-stage PwP since they have greater motor control and mental acuity. We acknowledge that it is more challenging to assess PD severity from the subtle and episodic symptoms in early-stage PwP. However, previous research [31] has successfully measured PD finger-tapping severity from video data collected remotely through a web-based, self-directed platform.

Furthermore, IoT sensor-based methods with near-perfect accuracy can be excellent tools for screening PwP with proper calibration mechanisms. It requires the acquisition of the sensors and the willingness of the participants to wear them. These sensors could also cost additional money and resources to set up, especially in the developing world. While IoT sensors are a completely reasonable way to screen PD, in our work, we attempt a more scalable and accessible method that can be deployed around the globe in both remote and in-person settings. As explained in Sec. 6.2, our framework can be extended to track PD longitudinally or assess other neurological disorders.

7.6 Demographics and Privacy-preference Information

In our study, we did not gather data regarding technical proficiency, socioeconomic status, various PD symptoms (motor, cognitive, memory-related, etc.), the severity of PD progression, the time elapsed since diagnosis, and similar factors. Additionally, our dataset solely originates from the US and comprises an English-speaking population, with 80 out of 91 (87.91%) individuals being of White ethnicity. Furthermore, we did not capture individuals' perspectives on sharing potentially sensitive information with external entities. In forthcoming iterations of our study, we intend to include all these details to enhance the context for interpreting our findings.

8 FUTURE WORK

8.1 Integrating Automated Quality Control

Upon inspecting the recorded videos, we discovered that the platform will benefit greatly from an automated quality control tool. Some recurring problems that are present in the videos (Fig 11) include:

(1) The participant's face was not in the center of the frame.
(2) Multiple people besides the participant are present in the video frame.
(3) The participants were reading out instructions during the recording.
(4) Bad lighting conditions in the filming environment.
(5) Hands not completely visible in extending arm task.

Fig 11 shows some of the problematic issues mentioned above. These issues highlight the necessity for a real-time validation pipeline that can provide feedback on video quality. We can disable the recording button until the framework has validated the filming environment. In recent years, numerous open-source tools such as mediapipe [35] and tensorflow.js [47] have enabled near real-time execution of facial recognition and pose

estimation. In our future endeavors, we intend to integrate these tools to assess the filming environment and provide real-time feedback to guide users during task performance. If data-quality is not adequately high, we can prompt the user to re-record the video.

8.2 Tailoring Resources to Each User

In the current iteration of our framework, we provide the same sets of resources to all the participants, irrespective of their unique circumstances. However, it would be prudent to customize the "Help & Survey" section based on users' results and demographics. For instance, for a participant at high risk of PD, we can provide more comprehensive information regarding PD, its symptoms, and the significance of early detection and treatment. Conversely, for a user at low risk of PD, we should emphasize the importance of maintaining a healthy lifestyle to minimize the risk of developing PD. In future iterations, we plan to collaborate with neurologists to determine the most appropriate resources for different user groups and integrate them into our framework.

8.3 Mitigating Risks of Misinformation in Chatbot

We integrated GPT-3 as the backbone of our chatbot. However, the large language models have a very well-known tendency to hallucinate – providing false information confidently – especially for questions requiring detailed and specific knowledge [4]. For example, when a participant asks for the phone numbers of neurologists in their area, the chatbot may generate specific phone numbers, website links, and addresses that are frequently nonexistent. To address this issue, we provided the GPT-3.5 API with detailed prompts and guidelines so that it refrains from engaging in off-topic conversation and provides preset website links for questions related to medication, neurologists, support groups, etc. However, despite extensive prompt tuning, we cannot guarantee that the chatbot will never generate harmful or misleading responses. Fortunately, researchers are making extensive efforts to mitigate the hallucination problem to make the models more reliable and robust [33]. Therefore, it is imperative that we continue to integrate these more reliable models and build appropriate guardrails simultaneously to ensure the real-world applicability of our framework.

8.4 Extending Our Lessons to Other Diseases

Although our framework was originally designed for Parkinson's disease, many of the lessons learned can be applied to developing similar frameworks for other diseases by taking the following measures.

8.4.1 Recording Environment. While we initially selected tasks based on MDS-UPDRS [25] and PARK framework [34], not all tasks were suitable for our use case. Specifically, the extended arm tasks presented significant challenges to users as they had to move further away from the laptop to ensure full visibility of their hands in the recording. In contrast, all other tasks could be completed without adjusting the initial position. Additionally, the extended arm task, intended to capture slight tremors in the patient's hands, may not be effectively captured by a laptop camera depending on the webcam's condition. These considerations are vital when designing similar frameworks for other diseases.

For instance, observing gait patterns in patients' walks is a standard task for assessing Ataxia [11]. However, incorporating a walking task would pose significant challenges as users would need to move within the webcam's view and some may be exposed to the risk of falling when completing the task without supervision. Recordings collected under these conditions might not be suitable for fully observing the user's gait patterns. Therefore, when designing digital frameworks for disease screening, it is critical to consider both the clinical validity of the tasks and the user's recording environment.

8.4.2 Selecting Intuitive Tasks. Another crucial consideration is to ensure that both the tasks and instructions are intuitive. During our study, as mentioned in Section 6, users found the distinction between anger and disgust

facial expressions to be ambiguous. This highlights the importance of selecting tasks that are clear and easily distinguishable.

A prime example of an intuitive task is the smiling expression task. Smiling is not only inherently intuitive, but it is also a universally recognized facial expression that can transcend cultural and demographic boundaries. It serves as an excellent benchmark for tasks that can be easily understood and performed accurately.

When selecting tasks for different diseases, it is paramount to carefully evaluate their intuitiveness and ensure that they do not lead to confusion or overlap with other tasks. By focusing on tasks that are easily comprehensible and distinct, we can enhance the effectiveness and reliability of the framework across diverse populations and disease contexts.

8.4.3 Designing User Interfaces Tailored to User Demographics. When designing user interfaces for a digital screening tool, it is crucial to consider the demographics of the target audience. In the case of Parkinson's disease, the majority of patients are over 50 years old. Therefore, additional considerations need to be in place to ensure that the framework is easy to navigate and the instructions are intuitive for this demographic. During the development of our framework, we enforced the following guidelines:

(1) Use large font sizes throughout the website.
(2) Provide explicit navigation instructions (e.g., "Click the 'Next' button to proceed") on every page.
(3) Ensure that links are clickable elements, eliminating the need for users to manually paste URLs into their browsers.
(4) Provide the opportunity to rewatch the instruction videos multiple times and re-record the task videos.

By following these guidelines, we ensured that the website has intuitive and user-friendly interfaces and controls for all age groups. Since the elderly demographic is often more susceptible to diseases, these guidelines and approaches can be extended to designing frameworks for other use cases as well.

8.5 Novel Application in Tracking PD Progression

Through our discussions with neurologists, we learned that our framework could also be deployed for a separate use case: tracking an individual's PD progression over time. Currently, PwP sees neurologists episodically (perhaps once every six months). Consequently, specialists only have a small, infrequent window through which to assess disease progression. Furthermore, PD symptoms may be episodic, which could mislead neurologists. Successful treatment of PD requires monitoring the disease continuously and fine-tuning regimens appropriately. We are currently adapting our tool to track PD progression. Specifically, we envision that users would complete framework tasks regularly over an extended duration to see how their results change over time; this data could be an important indicator of disease progression.

9 CONCLUSION

In this paper, we present a user-centered framework that enables anyone, anywhere, with access to a computer with a webcam and microphone to complete a set of neurological tests and receive a mock screening of their PD risk. Our framework also provides access to a chatbot driven by GPT, a neurologists' directory, and a list of PD prevention advice. To validate it, we launched a study enrolling 48 PwP and 43 controls and assessed their reception to it via both multiple-choice and open-ended questions. Of the 91 participants, 80.85% (standard deviation ±8.92%) of them either agreed or strongly agreed that they were in favor of this framework. The generated knowledge presented in this paper will be incorporated into subsequent iterations to implement: real screening after integrating machine learning models; automated quality control; tailoring resources to users based on their demographics and screening result, etc. As we see a further proliferation of digital tools in healthcare,

we hope that the design and ethical considerations, and the generated knowledge described in this paper will serve as a helpful reference.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Award IIS1750380, the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number P50NS108676, the Gordon and Betty Moore Foundation, and the Google Faculty Award.

REFERENCES

[1] Eman Abukmail, Mina Bakhit, Chris Del Mar, and Tammy Hoffmann. 2021. Effect of different visual presentations on the comprehension of prognostic information: a systematic review. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 1–10.

[2] Mohammad Rafayet Ali, Javier Hernandez, E Ray Dorsey, Ehsan Hoque, and Daniel McDuff. 2020. Spatio-temporal attention and magnification for classification of Parkinson's disease from videos collected via the internet. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 207–214.

[3] Mohammad Rafayet Ali, Taylan Sen, Qianyi Li, Raina Langevin, Taylor Myers, E Ray Dorsey, Saloni Sharma, and Ehsan Hoque. 2021. Analyzing head pose in remotely collected videos of people with Parkinson's disease. *ACM Transactions on Computing for Healthcare* 2, 4 (2021), 1–13.

[4] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[5] Wejdan L Alyoubi, Wafaa M Shalash, and Maysoon F Abulkhair. 2020. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked* 20 (2020), 100377.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[7] Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. 2022. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. *Frontiers in Digital Health* 4 (2022), 847991.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[9] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Ellen Buckley, Claudia Mazzà, and Alisdair McNeill. 2018. A systematic review of the gait characteristics associated with Cerebellar Ataxia. *Gait & posture* 60 (2018), 154–163.

[12] Benilda Eleonor V Comendador, Bien Michael B Francisco, Jefferson S Medenilla, and Sharleen Mae. 2015. Pharmabot: a pediatric generic medicine consultant chatbot. *Journal of Automation and Control Engineering* 3, 2 (2015).

[13] Norah L Crossnohere, Mohamed Elsaid, Jonathan Paskett, Seuli Bose-Brill, and John FP Bridges. 2022. Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks. *Journal of Medical Internet Research* 24, 8 (2022), e36823.

[14] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health Surveill* 6, 2 (9 Apr 2020), e18444. https://doi.org/10.2196/18444

[15] Mengyao Cui et al. 2020. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance* 1, 1 (2020), 5–8.

[16] Jari Dahmen, M Kayaalp, Matthieu Ollivier, Ayoosh Pareek, Michael T Hirschmann, Jon Karlsson, and Philipp W Winkler. 2023. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. , 3 pages.

[17] Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. 2023. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence* 6 (2023), 1169595.

[18] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer.

[19] E Dorsey, Larsson Omberg, Emma Waddell, Jamie L Adams, Roy Adams, Mohammad Rafayet Ali, Katherine Amodeo, Abigail Arky, Erika F Augustine, Karthik Dinesh, et al. 2020. Deep phenotyping of Parkinson's disease. *Journal of Parkinson's Disease* 10, 3 (2020), 855–873.

[20] E Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. 2018. The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's disease* 8, s1 (2018), S3–S8.

[21] E Ray Dorsey and Allison W Willis. 2013. Caring for the majority. *Movement disorders: official journal of the Movement Disorder Society* 28, 3 (2013), 261–262.
[22] Ray Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. 2020. *Ending Parkinson's disease: a prescription for action.* Hachette UK.
[23] CHAI (The Center for Human-Compatible AI). 2021. The Blueprint for Trustworthy AI in Healthcare. (2021). https://chai.berkeley.edu/blueprint-trustworthy-ai-healthcare/
[24] Rocio Garcia-Retamero, Edward T Cokely, and Ulrich Hoffrage. 2015. Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Frontiers in Psychology* 6 (2015), 932.
[25] CG Goetz, W Poewe, B Dubois, A Schrag, MB Stern, AE Lang, et al. 2008. MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). *Available from the International Parkinson and Movement Disorder Society website: https://www. movementdisorders. org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS. htm* (2008).
[26] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society* 23, 15 (2008), 2129–2170.
[27] Abid Haleem, Mohd Javaid, and Ravi Pratap Singh. 2022. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations* 2, 4 (2022), 100089.
[28] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.
[29] Kurtis G Haut, Adira Blumenthal, Sarah Atterbury, Xiaofei Zhoul, Wasifur Rahman, Emanuela Natali, M Rafayet Ali, and Ehsan Hoque. 2022. Assistive Video Filters for People with Parkinson's Disease to Remove Tremors and Adjust Voice. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
[30] Ashley M Hopkins, Jessica M Logan, Ganessan Kichenadasse, and Michael J Sorich. 2023. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectrum* 7, 2 (2023), pkad010.
[31] Md Saiful Islam, Wasifur Rahman, Abdelrahman Abdelkader, Phillip T Yang, Sangwu Lee, Jamie L Adams, Ruth B Schneider, E Dorsey, and Ehsan Hoque. 2023. Using AI to Measure Parkinson's Disease Severity at Home. *arXiv preprint arXiv:2303.17573* (2023).
[32] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882* (2022).
[33] Arvind Krishna Sridhar and Erik Visser. 2022. Improved Beam Search for Hallucination Mitigation in Abstractive Summarization. *arXiv e-prints* (2022), arXiv–2212.
[34] Raina Langevin, Mohammad Rafayet Ali, Taylan Sen, Christopher Snyder, Taylor Myers, E Ray Dorsey, and Mohammed Ehsan Hoque. 2019. The PARK framework for automated analysis of Parkinson's disease characteristics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–22.
[35] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
[36] Sarmad Maqsood, Robertas Damaševičius, and Rytis Maskeliūnas. 2022. TTCNN: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. *Applied Sciences* 12, 7 (2022), 3273.
[37] Bernard Marr. 2023. Revolutionizing Healthcare: The top 14 uses of CHATGPT in medicine and Wellness. https://www.forbes.com/sites/bernardmarr/2023/03/02/revolutionizing-healthcare-the-top-14-uses-of-chatgpt-in-medicine-and-wellness/?sh=4566df296e54
[38] Jeban Chandir Moses, Sasan Adibi, Nilmini Wickramasinghe, Lemai Nguyen, Maia Angelova, and Sheikh Mohammed Shariful Islam. 2022. Smartphone as a Disease Screening Tool: A Systematic Review. *Sensors* 22, 10 (2022), 3787.
[39] Michalis Mountantonakis and Yannis Tzitzikas. 2023. Using Multiple RDF Knowledge Graphs for Enriching ChatGPT Responses. *arXiv preprint arXiv:2304.05774* (2023).
[40] Nikhil Kishore Nayak, G Pooja, Ramya Ravi Kumar, M Spandana, and P Shobha. 2021. Health assistant bot. In *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2020, Volume 1*. Springer, 219–227.
[41] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *Knowledge and Systems Sciences: 18th International Symposium, KSS 2017, Bangkok, Thailand, November 17–19, 2017, Proceedings 18*. Springer, 38–52.
[42] Oded Nov, Nina Singh, and Devin M Mann. 2023. Putting ChatGPT's medical advice to the (Turing) test. *medRxiv* (2023), 2023–01.
[43] Wasifur Rahman, Sangwu Lee, Md Saiful Islam, Victor Nikhil Antony, Harshil Ratnu, Mohammad Rafayet Ali, Abdullah Al Mamun, Ellen Wagner, Stella Jensen-Roberts, Emma Waddell, et al. 2021. Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study. *Journal of Medical Internet Research* 23, 10 (2021), e26305.
[44] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Citeseer, 29–48.

[45] Ishith Seth, Aram Cox, Yi Xie, Gabriella Bulloch, David J Hunter-Smith, Warren M Rozen, and Richard Ross. 2023. Evaluating Chatbot Efficacy for Answering Frequently Asked Questions in Plastic Surgery: A ChatGPT Case Study Focused on Breast Augmentation. *Aesthetic Surgery Journal* (2023), sjad140.
[46] Krista G Sibley, Christine Girges, Ehsan Hoque, and Thomas Foltynie. 2021. Video-based analyses of Parkinson's disease severity: A brief review. *Journal of Parkinson's Disease* Preprint (2021), 1–11.
[47] Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Charles Nicholson, Nick Kreeger, Ping Yu, Shanqing Cai, Eric Nielsen, David Soegel, Stan Bileschi, et al. 2019. Tensorflow. js: Machine learning for the web and beyond. *Proceedings of Machine Learning and Systems* 1 (2019), 309–321.
[48] Andrej Thurzo, Martin Strunga, Renáta Urban, Jana Surovková, and Kelvin I Afrashtehfar. 2023. Impact of artificial intelligence on dental education: a review and guide for curriculum update. *Education Sciences* 13, 2 (2023), 150.
[49] Cristina Trocin, Patrick Mikalef, Zacharoula Papamitsiou, and Kieran Conboy. 2021. Responsible AI for digital health: a synthesis and a research agenda. *Information Systems Frontiers* (2021), 1–19.
[50] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports* 11, 1 (2021), 3254.
[51] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. 2022. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature medicine* 28, 10 (2022), 2207–2215.

A CHATBOT PROMPT

'X' Framework is an online tool for screening for Parkinson's Disease using AI developed by researchers at the University of 'Y', led by Dr. 'Z'. A participant can go to the 'X' website and perform different tasks similar to the MDS-UPDRS tasks and get a prediction of the likelihood of having Parkinsonian symptoms. The prediction from 'X' framework is not an official diagnosis of Parkinson's Disease. Any official diagnosis must be made by a neurologist.

The 'X' framework analyzes 3 types of tasks: Speech, Facial expression, and Motor, and predicts the participant's Parkinsonian symptoms in the following categories using machine learning models: Normal, Slight, Mild, Moderate, and Severe. Whereas "Normal" indicates a healthy individual and "Severe" means a high chance of having Parkinson's Symptoms.

You are 'X'-Bot. Your job is to help participants understand their scores in the 'X' framework, and help them learn more about Parkinson's disease. Do not answer specific questions such as medication, or information about a particular clinic or neurologist, instead refer them to the following links,

- Neurologist:https://www.google.com/maps/search/?api=1&query=neurologist
- Support groups: https://www.google.com/maps/search/?api=1&query=parkinson's+support+group
- Physical therapist: https://www.google.com/maps/search/?api=1&query=parkinsonâĂŹs+physical+therapist
- Medications: https://www.parkinson.org/living-with-parkinsons/treatment/prescription-medications

Do not provide any information that could be harmful, and do not engage in any off-topic conversation. Reply to the participant who received the following scores from 'X' framework,

- Speech symptoms: 'A'
- Facial symptoms: 'B'
- Motor symptoms: 'C'

The reports of the 'X' framework are estimations with machine learning, and may not be entirely correct.

B EXTRA FIGURES

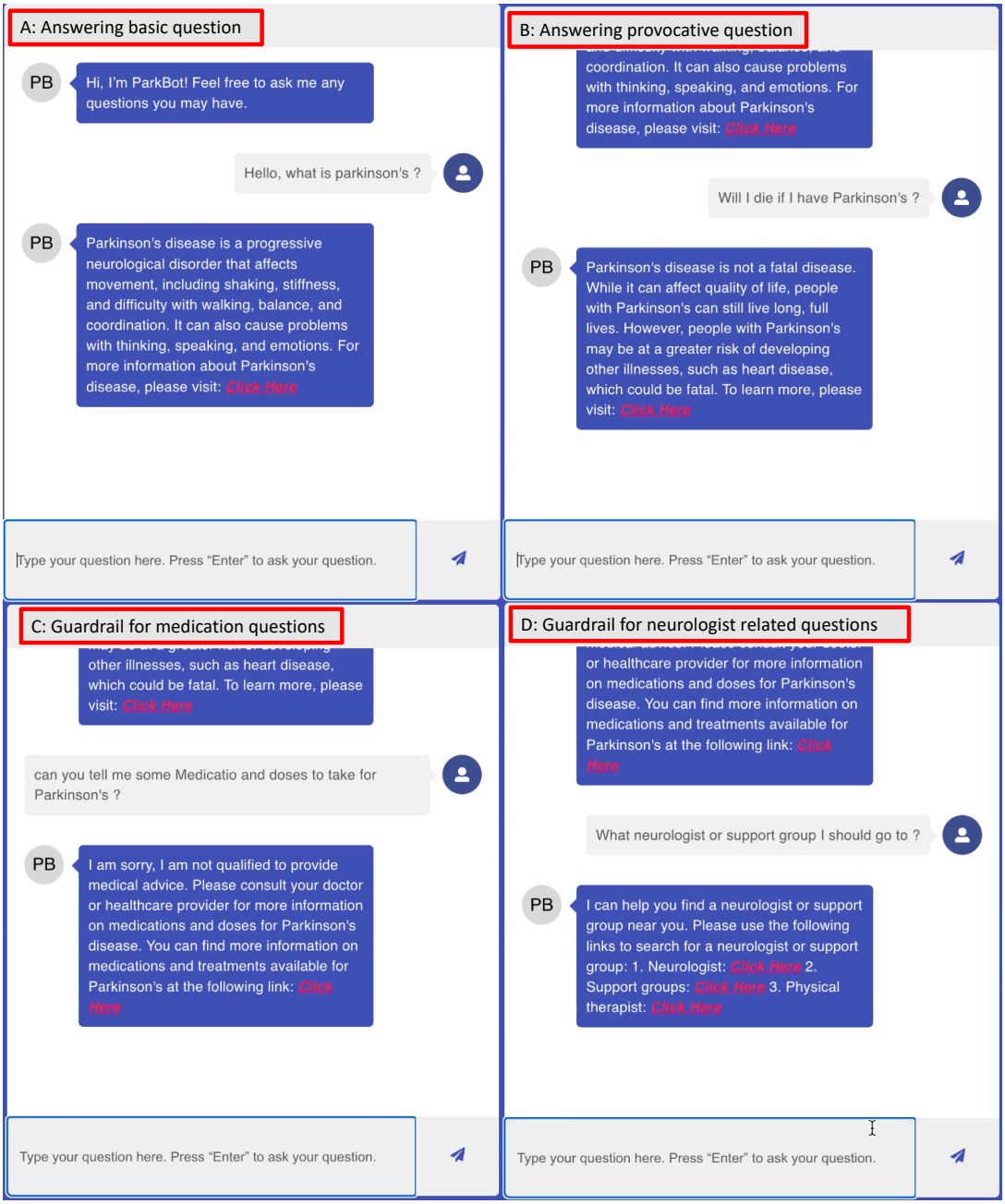

Fig. 4. Four examples of chatbot interaction: **A:** answering basic questions like, **B:** answering provocative questions, **C:** guardrails for medication questions, **D:** guardrails for neurologist related questions.

.

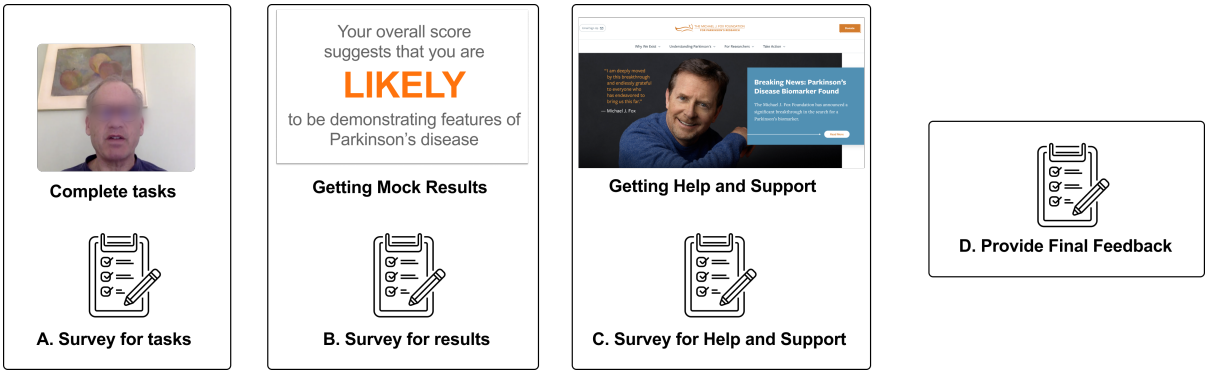

Fig. 5. Our validation study was conducted in four phases. At each phase, participants answer a set of multiple-choice and/or open-ended questions. Table 2 provides more information about the survey questions in each phase. **A:** Participants complete eight tasks (3). **B:** Then, participants get mock results(3). **C:** Next, the participants receive a set of resources – chatbot, neurologists directory, lifestyle advice on diet, exercise, sleep habit, etc. – as help & support(3). **D:** Then the participants complete ten questions from the System Usability Scale (SUS) to measure the usability of our framework.

.

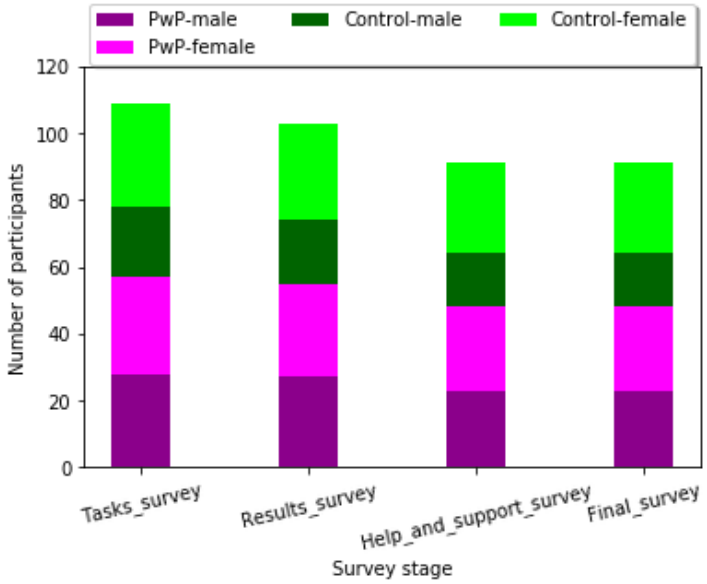

Fig. 6. The number of participants who completed each survey stage. 83.49% (91/109) of our participants who started the study finished it, indicating that our study was well-designed and capable of retaining a significant amount of participants.

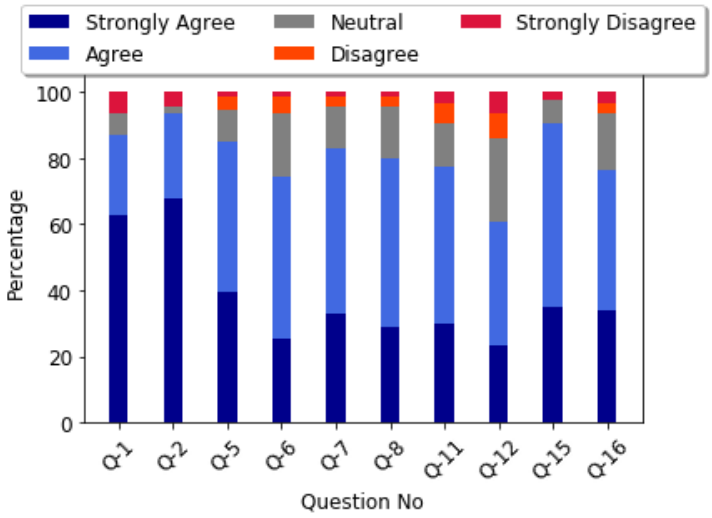

Fig. 7. Percentages of participants who responded Strongly Agree, Agree, Neutral, Disagree and Strongly Disagree to all the **multiple-choice questions** across **first three** study phases. Because the number of participants in each of the phases differed, we only use data from participants who completed the entire study (n = 91). The median and mean percentage of participants responding either Agree or Strongly Agree to the MCQs were 81.38% and 80.85% (standard dev ± 8.92) respectively.

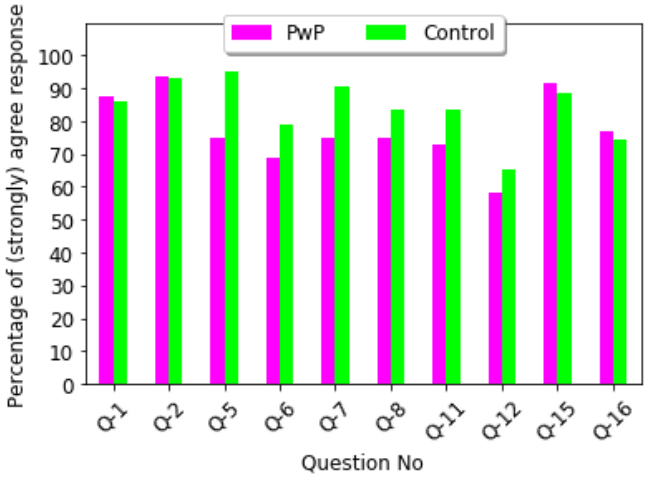

Fig. 8. Percentages of participants in PwP and Control who responded Strongly Agree or Agree to all the **multiple-choice questions** across **first three** study phases. Because the number of participants in each of the phases differed, we only used data from participants who completed the entire study. The median values of Agree or Strongly Agree percentage responses for the PwP and control groups were 75% and 84.88% respectively. Excluding Q-5 (p=0.007) and Q-7 (p=0.049), there isn't a statistically significant difference ($p < 0.05$) in the proportion of (Strongly) agree responses from PwP and the Control group (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$).

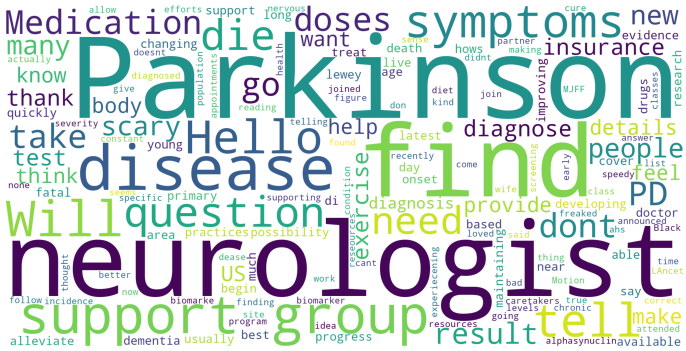

Fig. 9. Word cloud of user messages in the chat history with the chatbot.

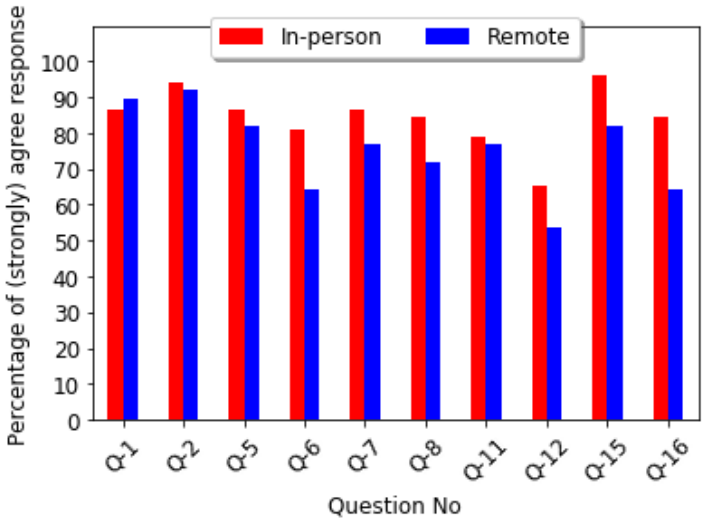

Fig. 10. Percentages of participants participating in-person and remotely who responded Strongly Agree or Agree to all the **multiple-choice questions** across **first three** study phases. Because the number of participants in each of the phases differed, we only use data from participants who completed the entire study. The median values of Agree or Strongly Agree percentage responses for the In-person and Remote groups were 85.58% and 76.92% respectively. However, excluding Q-15 ($p=0.025$) and Q-16 ($p=0.023$), there isn't a statistically significant difference ($p < 0.05$) in the proportion of (Strongly) agree responses from In-person and Remote groups (measured through two sample Z-tests for proportions with the level of significance, $\alpha = 0.05$).

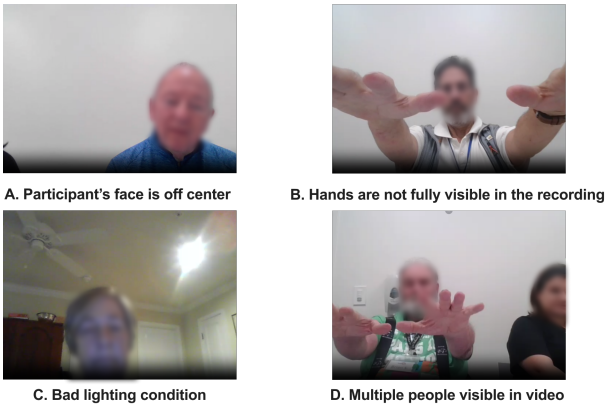

Fig. 11. Most of the participants were able to finish the tasks without any issues, but there were certainly cases where the videos has sub-par quality. The figure above presents four problematic cases we've found during our analysis.

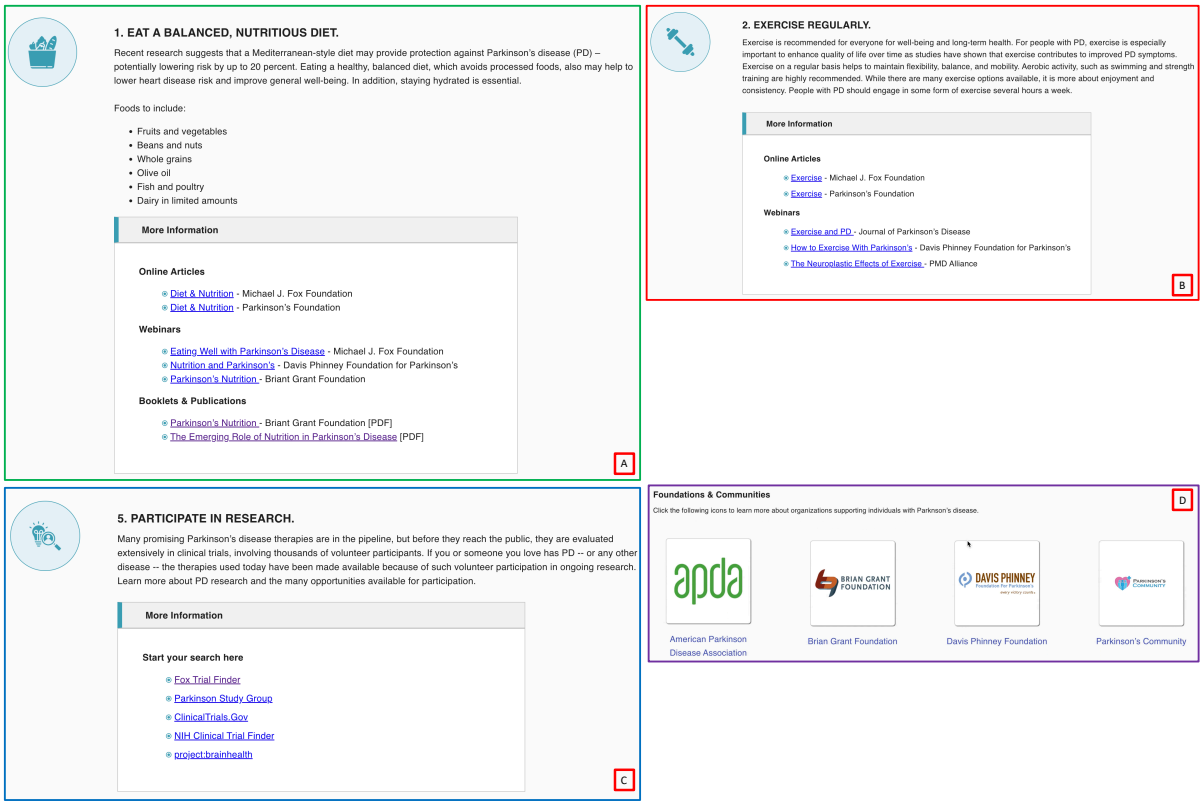

Fig. 12. Some of the resources provided to the participants.

.