

A Conversational Agent for Early Detection of Neurotoxic Effects of Medications through Automated Intensive Observation

Serguei Pakhomov[†], Jacob Solinsky, Martin Michalowski, Veronika Bachanova

*University of Minnesota,
Minneapolis, MN 55108, USA*

[†]*E-mail: pakh0002@umn.edu*

We present a fully automated AI-based system for intensive monitoring of cognitive symptoms of neurotoxicity that frequently appear as a result of immunotherapy of hematologic malignancies. Early manifestations of these symptoms are evident in the patient's speech in the form of mild aphasia and confusion and can be detected and effectively treated prior to onset of more serious and potentially life-threatening impairment. We have developed the Automated Neural Nursing Assistant (ANNA) system designed to conduct a brief cognitive assessment several times per day over the telephone for 5-14 days following infusion of the immunotherapy medication. ANNA uses a conversational agent based on a large language model to elicit spontaneous speech in a semi-structured dialogue, followed by a series of brief language-based neurocognitive tests. In this paper we share ANNA's design and implementation, results of a pilot functional evaluation study, and discuss technical and logistic challenges facing the introduction of this type of technology in clinical practice. A large-scale clinical evaluation of ANNA will be conducted in an observational study of patients undergoing immunotherapy at the University of Minnesota Masonic Cancer Center starting in the Fall 2023.

Keywords: Large language models, artificial intelligence, speech, language, immunotherapy, Immune effector cell-associated neurotoxicity syndrome

1. Introduction

Immune effector cell-associated neurotoxicity syndrome (ICANS) represents a unique complication of immune effector therapy particularly in patients treated with chimeric antigen receptor T-cell therapy (CAR-T) cells for hematologic malignancies. ICANS incidence varies from 40-60% depending on specific CAR-T product and grading using a 4-point scale, with 1 being the mild manifestation and 4 the most severe. ICANS usually presents 3-5 days after CAR-T infusion and about 20% of events present at grade 3 or higher. The clinical centers administering approved CAR-T therapies have to comply with the Risk Evaluation and Mitigation Strategies (REMS) mandated by the Food and Drug Administration (FDA). These include monitoring and prompt treatment of ICANS symptoms. The purpose of ICANS monitoring and detection after the CAR-T infusion and prompt treatment is to halt ICANS progression and minimize the risk of brain edema/herniation, the most feared sequelae of ICANS resulting in severe cognitive impairment, coma, ICU stay, intubation, and, in rare cases, death.¹⁻³

© 2023 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Clinical manifestations of ICANS typically begin with word-finding difficulty, headaches, confusion, dysphasia, aphasia, impaired fine motor skills resulting in agraphia, and somnolence⁴ and, if untreated, can progress to the more severe sequelae. Expressive aphasia has been found to be the most specific symptom of ICANS. It starts as impaired ability to name objects, paraphasia errors, hesitant speech, and verbal perseveration, which can then proceed to global aphasia (inability to speak or respond to commands) with increasing ICANS severity.⁴ In fact, initial expressive aphasia is highly prevalent (86%) in patients that then go on to develop severe neurotoxicity.⁵ Low-grade ICANS is managed predominantly by supportive care or low dose dexamethasone, whereas severe ICANS is usually treated with high doses of corticosteroids and anakinra which can partially block the cascade of inflammation leading to pathology.⁶ Recently emerging clinical evidence suggests that early intervention with a short course of corticosteroids such as dexamethasone in patients with low-grade ICANS can resolve these symptoms completely and thereby prevent progression to more severe ICANS.⁷ However, administration of corticosteroids as prophylaxis of ICANS in all patients undergoing CAR-T therapy is not desirable as corticosteroids may have a negative impact on the effectiveness of CAR-T therapy itself, have short and long-term side-effects, increase risk of infections and therefore lower dose and short course is desirable.⁸

The existing methods for detecting neurotoxicity of immunotherapy (as described in the National Comprehensive Cancer Network (NCCN) guidelines) consists of administering brief cognitive assessment tools such as the Immune Effector Cell-Associated Encephalopathy (ICE) Assessment Tool or the CAR-T Cell Toxicity Tool (CARTOX-10). Both are loosely based on the Mini-Mental State Examination (MMSE) originally developed for the diagnosis of dementia and include several brief cognitive instruments. The CARTOX-10 consists of the following 4 categories: Orientation: orientation to year, month, city, hospital, president of country of residence (5 pts); Naming: ability to name 3 objects (e.g., point to clock, pen, button) (3 pts); Writing: ability to write a standard sentence (e.g., “Our national bird is the bald eagle”) (1 pt); Attention: ability to count backwards from 100 by 10 (1 pt). ICE adds one more category to the CARTOX-10 instrument: Following commands: ability to follow simple commands (e.g., “Close your eyes and stick out your tongue”) (1 pt).

These tools are widely used for screening for ICANS, are brief and easy to use at bedside, and are highly specific for ICANS but lack scientifically rigorous evaluation. These tools inherited low sensitivity from the MMSE on which they were based, as evidence from practice suggests that patients in the early stages of ICANS may pass the ICE assessment (especially if they are able to memorize it due to its frequent administration) while displaying some of the more subtle ICANS symptoms.⁹ Another major drawback of the existing screening tools is that while these paper-and-pencil tests are not particularly difficult to administer and score, their administration requires a qualified healthcare provider and is time consuming. Since post CAR-T therapy follow-up requires intensive daily monitoring usually for up to 14 days, that introduces a significant burden on clinical personnel and healthcare resources. This routine practice limits the frequency and depth to which patients can be feasibly monitored with ICE and, consequently, may lead to missing the onset of early symptoms in between assessments. Using technology to help in administering and facilitating more frequent follow-up of patients

would be a significant advance in assuring safer CAR-T therapies by enabling earlier detection of more subtle symptoms. There is also an increasing trend to administer CAR-T therapy in the outpatient setting. At-home monitoring of ICANS symptoms is highly desirable as it offers the potentially more timely intervention while providing patients with more comfort and convenience.

Early detection of ICANS would allow using lower doses of corticosteroids but would also require intensive monitoring of cognitive function (e.g., 3 times per day vs. the typical once per day frequency). Infrequent monitoring for ICANS (once a day or less) is likely to miss the early onset of subtle symptoms, as demonstrated by a study of 133 patients undergoing CAR-T therapy.¹⁰ Fifty-one of these patients developed ICANS and 27 of the 51 patients (53%) presented already with Grade ≥ 2 ICANS as the initial diagnosis. According to the ASTCT Consensus Grading guidelines, Grade 2 ICANS is diagnosed when the patient scores in the range 3-6 (out of 10 possible points). In practical terms, to score 3-6 on the ICE test, the patient would have to be significantly impaired (i.e., unable to tell what year, month it is, which city or hospital they are in, who the president is, and/or name three basic objects). The fact that over half of the patients with ICANS are initially diagnosed with Grade 2 or higher, combined with the fact that ICANS can develop in a matter of hours, indicates high likelihood that milder symptoms were present earlier but were missed either due to poor sensitivity of ICE, its relatively infrequent administration, or both.

Limitations of the standard-of-care approaches to ICANS detection combined with the availability of highly effective therapy to prevent its further progression⁷ create the urgent need for a validated, low provider burden, and well-tolerated by patients solution for early identification of neurotoxicity. Deploying such a solution will potentially result in preventing an estimated 40-70% of cancer patients who are at risk of ICANS from severe and potentially debilitating symptoms. An effective solution will also reduce the total dose and duration of steroids, mitigate the steroid effect on CAR-T function and response, and can potentially improve CAR-T outcomes, enable easier access to CAR-T for older people, and facilitate outpatient administration and management after CAR-T therapy.

In this paper, we provide a description of the design and implementation of an Automated Neural Nursing Assistant (ANNA) system designed to address the limitations of the standard-of-care approaches by automating the administration and analysis of speech-based neurocognitive tests^a. We also discuss the challenges specific to this particular clinical use case of intensive monitoring for cognitive changes associated with neurotoxic effects of immunotherapy, as well as other emerging areas where such intensive monitoring may be needed. We also report on the results of a small preliminary functional evaluation study designed to evaluate user experience with the system and collect feedback to determine areas for improvement prior to conducting a clinical study scheduled to begin in the Fall of 2023.

^aA live demo version of ANNA has been presented at the 2023 Interspeech symposium and is currently available at +1 (612)-682-6292. Note: the phone number may change over time - to obtain the current number for the demo, please contact the authors

2. System Description

ANNA consists of a multi-platform app (iOS, Android, telephony) that administers neurocognitive tests, collects voice responses, and securely uploads them to a web service that stores the audio and automatically scores the tests. The implementation described in this paper operates via the telephone interface. To make the conversation as natural as possible, the system is implemented to work in full-duplex audio mode in which both the patient and the system can speak at the same time without the need for the patient to signal the end of utterances by pressing a button.

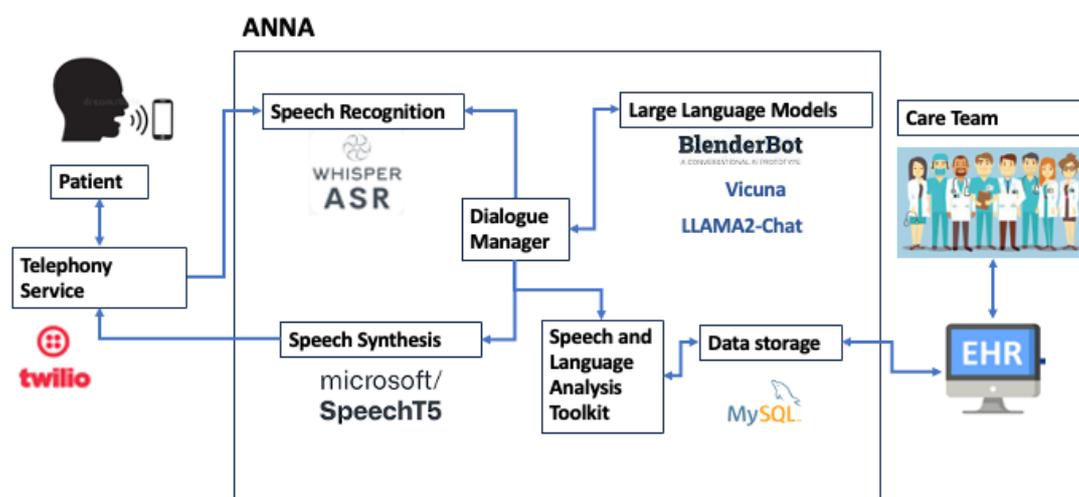


Fig. 1: Illustration ANNA system architecture and data flow.

ANNA's architecture illustrated in Figure 1 consists of two independent components, a dialogue manager and a cognitive assessment test battery. As the conversation manager goes through the script of the phone call, it transcribes the patient's speech, responds with synthesized speech, plays audio, and listens for pauses and cue words from the patients speech to allow ANNA to take turns in conversation in a natural manner. For speech transcription it employs OpenAI's state-of-the-art Whisper transcriber, which we have found to produce acceptable transcriptions of audio recordings from even the lowest end consumer phones. For speech generation we use the the pre-trained SpeechT5 model.¹¹ Twilio^b currently provides telephony services to ANNA, however, the Dialogue Manager can easily be reconfigured to accept input from and produce input for other audio recording and playback devices, allowing us to reuse it in our group's other voice application projects. The Dialogue Manager currently consists of a set of rules that are designed to walk the patient through the process of

^b<https://www.twilio.com/>

participating in the cognitive assessment. The conversation manager can either read directly from a script, which is how it conducts the word-list cognitive assessments or it can prompt a Large Language Model (LLM) using the patient’s last utterance to generate a response. For this purpose we currently use the blenderbot-400M-distill¹² model and insert the responses to the input received from the patient at each conversation turn. Blenderbot is pre-configured to understand dialog and uses no prompt, responding directly to the patient’s utterance. We also continue to experiment with other pre-trained LLMs (limited to those that can be used in a local HIPAA-compliant environment) including the Vicuna Chat¹³ and, most recently, Llama2 Chat¹⁴ models.

We developed ANNA as an easily deployed set of Docker images which can be deployed within an on-site server when provided with a phone number, web address, and a GPU with at least 24Gb of VRAM (e.g., NVIDIA RTX 3090 Ti). The current demo implementation is running on a server with two NVIDIA RTX 4090 cards. We have also constructed an alternative implementation of ANNA which does not use Docker images for any components which require access to a GPU, as the fully containerized application can have difficulty accessing the GPU in some environments.

2.1. Spontaneous Speech and Language Elicitation

We programmed ANNA to make a phone call to the patient’s phone (smartphone or landline) and administer the following tasks: a brief conversation with a conversational agent based on a LLM that asks the patient to describe how they are feeling and conducts a brief conversation on one of a set of pre-defined topics such as favorite pastime, books, movies, etc. Topics are currently randomly drawn from a pre-defined list without replacement to alleviate practice effects.

2.2. Cognitive Testing: Word List Recall

The conversation is followed by a series of brief cognitive tasks including a word list learning task in which the patient is presented with a list of 6 words and is asked to recall as many of these words as the patient can immediately after the presentation (immediate recall) and a few minutes later (delayed recall). The word list recall task is vulnerable to practice effects in serial testing.¹⁵ Practice effects can mask subtle cognitive changes due to early stages of ICANS; therefore, we developed a mechanism for generating multiple alternative lists of 6 words to minimize the effects of repeated test administration. To ensure that the lists of words are roughly equivalent across multiple presentations, we developed an approach for automatically generating lists of words that are equivalent in their lexical properties of frequency, concreteness, and imageability using the MRC Psycholinguistic database.¹⁶

2.3. Cognitive Testing: Verbal Fluency

Two verbal fluency tests are administered between the immediate and delayed recall tasks. The verbal fluency tests consist of a category fluency test in which the patient is asked to name as many animals as they can think of in 30 seconds, followed by a letter fluency test asking to

name as many words beginning with the letter “F” as they can think of also in 30 seconds. The verbal fluency task also suffers from practice effects; however, prior work of other researchers and our own preliminary data show that in these generative tasks the practice effects are small and plateau after several presentations in individuals with cognitive impairment.¹⁶ The rationale for selecting verbal fluency and list learning tasks rests on the evidence that they are particularly sensitive to a broad-spectrum of cognitive impairment effects caused by a wide variety of acute and chronic conditions including effects of medications,¹⁷ are quick to administer, and lend themselves well to automation.

We selected the abbreviated versions of the list learning and verbal fluency tests to make them less burdensome for patients undergoing cancer treatment. The abbreviated version have been shown to have similar psychometric properties to their full counterparts (10 words for the list learning and 60 seconds for the verbal fluency tests).^{18,19}

2.4. Speech and Language Analysis

The speech collected with ANNA is first subjected to automatic speech recognition to produce a verbatim transcript of everything the patient said during the interaction with the system. The current implementation of ANNA relies on the pre-trained Whisper neural transformer model (large-v1).²⁰ The transcribed speech is analyzed to extract the following language characteristics: syntactic complexity and language model perplexity. Syntactic complexity is measured using technology we previously developed to characterize language changes in patients with dementia.²¹ Measures of syntactic complexity include the mean number of clauses, various measures of the depth of syntactic trees obtained from a syntactic constituency parser, and the mean syntactic dependency length obtained from a dependency parser. Perplexity is a measure of how many different equally most probable words can follow any given word based on probabilities obtained from a probabilistic or a neural language model. High mean perplexity computed over an utterance that did not participate in training the model indicates a poor fit between a language model and the text of the utterance. This measure has been shown to be useful for distinguishing between speech of individuals with probable Alzheimer’s Disease and healthy controls.²² Both the syntactic complexity and the language model perplexity have been included in an attempt to capture early signs of confusion and changes in language patterns that have been noted in patients starting to develop ICANS. In addition to the language characteristics described above, we also extract the following paralinguistic speech characteristics: initial pause duration prior to onset of speech, mean pause duration, perseveration and hesitation density. These characteristics have also been noted (anecdotally) by oncology treatment teams as being observed in early stages of ICANS.

2.5. Motivation for Selecting Analytical Measures

Rigorous prior work by other investigators and by our group demonstrated that the category verbal fluency task (animal naming) discriminates between individuals with Alzheimer’s disease dementia and healthy controls with sensitivity of 0.88 and specificity of 0.96.²³ The list learning task has also been previously shown to have excellent psychometric properties for detecting mild (mean MMSE score = 22.1) memory impairment (sensitivity and specificity of

90%) and high test-retest reliability.²⁴ In our own prior work, we have demonstrated that using deep neural modeling of the spontaneous speech patterns produced during a picture description task is able to discriminate between Alzheimer’s disease dementia and controls with 87% accuracy.²² Our team has been engaged in validation studies in which we have demonstrated that neurocognitive tests of verbal fluency (semantic and phonemic) are highly sensitive to the neurotoxic effects of psychoactive medications such as lorazepam and topiramate,²⁵ as well as the effects of chronic traumatic encephalopathy,²⁶ effects of nicotine withdrawal²⁷ and neurodegeneration due to Alzheimer’s disease.²⁸ While ICANS is distinctly different from the slowly progressing symptoms in dementia and chronic traumatic encephalopathy, it is similar to the rapid (on the order of hours) changes in cognition observed as a result of acute effects of psychoactive medications and nicotine withdrawal.

2.6. Pilot Evaluation Study

We conducted a small functional evaluation study of ANNA to elicit initial feedback from healthy individuals that could inform any further changes in system design and help us debug the system. We asked 10 Amazon Mechanical Turk workers to place an anonymous call to ANNA, interact with the system, and respond to a brief survey shown in Figure 3.

- **ANNA:** Hello, this call will be recorded for research use. Because I am a robot, I can be slow sometimes. Please don't hang up if I take too long to respond or say things that don't make much sense. I am going to ask you a few questions. Please listen carefully and answer them in as much detail as you can. Are you ready?
- **CALLER:** Yes.
- **ANNA:** Okay. First, I would like to ask you how you are feeling. For example, have there been any changes in how you feel in the last few hours?
- **CALLER:** No, I think I feel great. I have been enjoying the weather and I think overall, day's been going great and no changes in how I am feeling.
- **ANNA:** That's great! I'm glad you're having a good day. What's the weather like where you are?
- **CALLER:** Right now it is 97 degrees Fahrenheit. It's sunny. We're supposed to get triple digit 100 degree weather probably for the next seven days. But I mean, besides that, I'm in a cool environment and trying to limit my time outside.
- **ANNA:** Thank you for sharing. Now, I would like to ask you to tell me as much as you can about your most memorable trip. Feel free to take your time in answering.
- **CALLER:** Okay, well I would say most memorable trip is, this is a long time ago, this is back in 1999, in December of 1999, took a trip to Miami, Florida.
- **ANNA:** That sounds like a lot of fun. I've never been to Florida. What did you do there?
- **CALLER:** Well, went to Disney World, went to the beach, went to Miami, went to Orlando.
- **ANNA:** Thank you. Now, I am going to ask you to remember a list of six words. You will hear six words being spoken to you. Please repeat each word aloud as it is being spoken to you. Later, I will ask you to recall all six words. Are you ready?
- **CALLER:** Yeah, that's what we – it's pretty much what we did. Yes.
- **ANNA:** Here is the list: turtle, spray, lion, yard, queen, land. Now repeat as many of these words as you remember and say next when you are done
- **CALLER:** Turtle. Prey. Lion. Yard. Plane. Land. Turtle. Prey. Land. Farm. Trying to think of the other ones.
- **ANNA:** Thank you. Now, I will give you a letter of the alphabet. I am going to ask you to name words that begin with that letter, as fast as you can. For example, if I give you the letter S, as in sam, you can say soft, smile, and so on. Do not use the same word with a different ending such as smiling, or smiles. Are you ready?
- **CALLER:** Yes.
- **ANNA:** Okay. Your letter is the letter F, as in foxtrot. Please name all the words that you can think of that begin with the letter F. You have thirty seconds. Please begin.
- **CALLER:** Field, final, fantasy, fortune, fox, frisbee,
- **ANNA:** Please stop. Now, I will give you a category. I am going to ask you to name as fast as you can all the things that belong to that category. For example, if I give you the category of articles of clothing, you can say shirt, or jacket, or pants, and so on. Are you ready?
- **CALLER:** Yes.
- **ANNA:** Okay. Your category is animals. Begin naming as many animals as you can think of. You have thirty seconds.
- **CALLER:** Go ahead.
- **ANNA:** Please stop. We are almost done. Just one last thing. A few minutes ago I read a list of six words to you. Please try to recall as many of these words as you can and say them aloud as you remember them. You have thirty seconds. Please begin.
- **CALLER:** Tiger. Field prey, turtle.
- **ANNA:** Thank you. This concludes our session. Until next time. Goodbye.

Fig. 2: Example of an interaction between an Amazon Mechanical Turk worker and ANNA. This example shows the actual unedited transcription of the caller’s voice with the Whisper transcriber.

3. Results

An example interaction between a functional evaluation study participants and ANNA is shown in Figure 2. This example shows a verbatim transcript of the interaction which illustrates the performance of all ANNA components including the automatic speech recognition and large language models. The quantitative results of the functional evaluation are summarized in Table 1.

The Duration column in Table 1 reflects the amount of time it took the evaluator to interact with ANNA and complete the evaluation survey. The mean duration for the 10 evaluators was 11 minutes. All evaluators had a 100% approval rating on the Amazon Mechanical Turk system (i.e., they were approved for payment for all human intelligence tasks that they performed in the past). All evaluators were able to get to the end of the interaction with ANNA successfully (Completed column in Table 1). The mean audibility rating was 4.1 (SD: 0.74), the sensibility of ANNA’s responses to evaluators was rated as 3.7 (SD: 0.95), and the latency of system responses was rated as 2.0 (0.92).

Table 1: Results of pilot functional evaluation.

Evaluator	Duration (sec.)	Completed Y/N	Audibility (1-5)	Sensibility (1-5)	Latency (1-5)	Comments
1	660	Y	5	3	1	–
2	669	Y	4	3	4	Good exprience
3	660	Y	5	3	1	–
4	557	Y	3	5	2	Make it a little bit faster in response.
5	784	Y	4	5	2	–
6	808	Y	4	3	3	It could be a bit more human-like, now it sounds too machine like.
7	782	Y	3	5	2	Just improving response time would be a huge upgrade.
8	744	Y	5	3	1	–
9	748	Y	4	3	2	GOOD
10	588	Y	4	4	2	It should be able to restate the instructions instead of just waiting on an affirmative to being.
Mean (SD)	700 (86.18)	Y	4.1 (0.74)	3.7 (0.95)	2.0 (0.92)	–

4. Discussion

The broad clinical need that ANNA is designed to address arises from the limitations of the healthcare system in which intensive monitoring of patients' cognitive function (multiple times per day) by a human healthcare professional is not feasible and is cost-prohibitive. Monitoring cognitive function is unlike monitoring of physiologic function in that the former requires symbol-mediated interaction, which is typically achieved through the use of language. Many years of research in language technology and artificial intelligence yielded a number of conversational agents designed for use in healthcare applications.²⁹ However, recent developments in speech and language technology and, in particular, the introduction of large language models such as ChatGPT and Whisper can potentially move these efforts to a new level by making these systems simpler and more accurate in recognizing the incoming speech and producing more natural and flexible responses.

Were you able to get through the entire script of the phonecall?

On a scale of 1-5, with 1 being slow and 5 being fast, how quick was the chatbot in responding in conversation?

On a scale of 1-5, with 1 being nonsense and 5 being sensible, how sensible were the things the chatbot said back to you?

On a scale of 1-5, with 1 being difficult and 5 being easy, how easy was it to hear the words that were spoken in the word memorization test?

Additional comments on how we can improve our chatbot system and its user experience

Fig. 3: Evaluation survey administered to Amazon Mechanical Turk workers.

The proposed automated cognitive assessment methods address the limitations of the existing manual methods by using a series of brief, validated and easy to administer neurocognitive tests that use speech as the input modality to detect expressive aphasia deficits - the most specific symptom of ICANS. The key innovative aspect of our approach is the use of AI technologies such as LLMs and automatic speech recognition based on deep learning to convert spoken responses to text that can subsequently be used to compute traditional scores as well as novel speech and language-based measures to further improve the sensitivity and specificity of these instruments. The use of AI, large language models and automatic speech recognition and synthesis, as well as scoring algorithms tailored to the neurocognitive tests at hand, is what sets us apart from other commercial and academic computerized neurocognitive testing

approaches. To the best of our knowledge none of the current computerized approaches to neurocognitive assessment use conversational AI technology to elicit speech from patients and to analyze the resulting speech for cognitive impairment due to immunotherapy with machine learning. Another innovation is that we use extensively validated and recognized neurocognitive tests in a novel, accessible, and fully automated way that can also enable at-home and/or remote monitoring for ICANS, which could improve the accessibility of immunotherapy in rural and other settings away from medical centers.

In addition to the immunotherapy used to treat certain types of cancer, other newly emerging therapies that leverage the immune system have been recently approved by the US Food and Drug Administration for treatment of Alzheimer’s disease. The most recent approval was granted in July 2023 to lecanemab, an immunotherapy agent that was demonstrated to remove Alzheimer’s disease biomarkers from the brain and significantly (albeit moderately) slow down the disease progression as compared to other treatments. However, serious side effects including brain edema in 12.6% of the participants in the active arm of the clinical trial of this medication were observed.³⁰ Therefore, similarly to ICANS, early detection and clinical management of these side-effects in the treatment of Alzheimer’s disease may potentially benefit from intensive cognitive monitoring. Another potential clinical application area for systems like ANNA is in automating the monitoring for post-operative delirium. Proactive monitoring for post-operative delirium and early intervention has been shown to shorten length of hospital stay and improve surgical outcomes.³¹

Pending successful demonstration of ANNA’s feasibility and validity for early detection of ICANS, as we move outside of the realm of research and into wide adoption of ANNA in clinical practice, ANNA is ready to be integrated into a wide variety of clinical settings as a laboratory service using already existing technology and informatics standards including the Health Level 7 (HL7 v2) and FHIR protocols to interface with EHR via the standard lab test results route. One of the challenging issues that we expect to face has to do with handling of critical values. Critical values or failure to do the ANNA assessment will need to be communicated to the care team verbally by phone also using a standard protocol (ISO 15189) for communicating critical lab values. To this end, ANNA would need to have an interface (voice or graphical) to enable the care team to configure the system for each individual patient. The configuration will need to include telephone numbers for the patient and the care team as well as some of the patients’ preferences (e.g., topics for the conversational part of ANNA’s assessments, do-not-call times, system voice and personality preferences).

4.1. *Limitations*

ANNA currently has several technical limitations. Due to the use of multiple large neural models, the response latency can vary between less than a second for short turns (e.g., confirmations) to 3-4 seconds for longer turns in which ANNA has to convert longer input utterances to text and then also generate a response text and synthesize it into a spoken utterance. The evaluators in the pilot study clearly noted this as something that should be improved. We found that neural text-to-speech generation is the biggest contributor to response latency; however, other modules can be optimized as well. We plan to reduce the response latency in

the production version of ANNA by a) switching to a faster version of the Whisper model^c which has been benchmarked to be about 5 times faster than the OpenAI version, and b) distributing the LLM and TTS models across multiple GPU cards.

Another potential limitation that ANNA inherits from the pre-trained large language models is the potential for going off-topic (a.k.a. "hallucinating") during the initial conversational part of the assessment. To minimize this potential risk, we limit the amount of text produced by the models in response to the user input to 1-2 utterances. In the near future, we also plan to implement a set of guardrails to prevent ANNA from responding in inappropriate or offensive manner^d. This limitation was not noted in the pilot study as the sensibility of ANNA's responses was rated as fairly high (mean 3.7 out of 5) and none of the 10 evaluators commented on any specific nonsensical responses.

The racial, cultural, gender and ethnic biases learned by large language models from training data is a major concern with applications of AI in medicine in general³² and is a potential concern in our application as well. Given the nature of the interactions between patients and ANNA and the focus on eliciting as much speech from patients as possible over as few conversational turns as possible, we do not anticipate any such biases to have a chance to manifest themselves in any discernible fashion to the patients. Nonetheless, since inherent bias in language models is a known issue and we plan to examine the data collected with ANNA for any signs of bias or unfairness and experiment with current de-biasing methods.

ANNA's use case also has a distinct strength with respect to one of the biggest known limitations of large language models - variable trustworthiness of the information they generate. The lack of confidence in the information provided by these models is currently one of the major barriers to their adoption for clinical applications as primary sources of clinical knowledge.³³ ANNA's clinical use case, however, does not rely on large language models for knowledge. We rely on these models only to support a chatbot application used to elicit speech from patients for subsequent analysis and not to inform either patients or clinicians. As such, ANNA currently represents one of the safest and most immediate ways of using large language models in a clinical context.

5. Next Steps

We have developed and submitted an observational study protocol to the University of Minnesota Institutional Review Board. In this prospective clinical study University of Minnesota Masonic Cancer Center patients undergoing CAR-T therapy for hematological cancers will be monitored for ICANS with ANNA concurrently with the standard of care ICE testing. The following primary endpoints will be evaluated: a) acceptability of the frequency of ANNA administration; b) quality and quantity of audio collected from patients; and c) naturalness and ease of interaction with automated ANNA assessments. As we test the central feasibility hypothesis, we will also seek to understand the reasons why ANNA administration may not have occurred (examples: unable, refused, too tired, ill, forgot, technical reason, app not

^c<https://github.com/guillaumekln/faster-whisper>

^d<https://github.com/NVIDIA/NeMo-Guardrails>

working, battery out, other). We will also evaluate ANNA's usability characteristics that are not central to its feasibility but may affect the feasibility indirectly such as naturalness of interactions with patients, convenience, and patients' perceptions of ease of use.

Prior to conducting the clinical study, we plan to address the system latency limitation pointed out by the pilot study evaluators as well as experiment with the more recently released chat models such as Llama2 to improve the sensibility of the initial conversations with the patient.

We also plan to enhance the language analysis of the conversations collected with ANNA by adding language coherence measures using a recently developed Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS) method that relies on a time-series analysis of coherence features computed using semantic relatedness between words in a given piece of discourse^e. The TARDIS approach has been used successfully to characterise disordered speech in patients with schizophrenia³⁴ and may prove to be useful for detecting possible thought disturbances caused by early ICANS.

One of our current concerns with using ANNA for intensive monitoring of cognitive changes in cancer patients is that even the abbreviated version of the cognitive tests we have currently implemented may present a burden for the patients who are likely to experience significant distress and fatigue as a result of therapy. Our ultimate goal in the forthcoming clinical study is to determine if we can reliably ascertain the onset of ICANS based entirely on the analysis of the brief conversation between ANNA and the patient. If we can successfully do so, then we would likely be able to dispense with the more formal word list learning and verbal fluency tests, which would make intensive monitoring much less burdensome for patients.

References

1. M. S. Topp, T. van Meerten, R. Houot, M. Minnema, N. Milpied, P. J. Lugtenburg, C. Thieblemont, M. Wermke, K. Song, I. Avivi, J. Kuruvilla, U. Dührsen, R. Chu, L. Zheng, V. Plaks, A. Kerber and M. J. Kersten, Earlier Steroid Use with Axicabtagene Ciloleucel (Axi-Cel) in Patients with Relapsed/Refractory Large B Cell Lymphoma (R/R LBCL), *Biology of Blood and Marrow Transplantation* **26**, p. S101 (March 2020).
2. S. S. Neelapu, F. L. Locke, N. L. Bartlett, L. J. Lekakis, D. B. Miklos, C. A. Jacobson, I. Braunschweig, O. O. Oluwole, T. Siddiqi, Y. Lin, J. M. Timmerman, P. J. Stiff, J. W. Friedberg, I. W. Flinn, A. Goy, B. T. Hill, M. R. Smith, A. Deol, U. Farooq, P. McSweeney, J. Munoz, I. Avivi, J. E. Castro, J. R. Westin, J. C. Chavez, A. Ghobadi, K. V. Komanduri, R. Levy, E. D. Jacobsen, T. E. Witzig, P. Reagan, A. Bot, J. Rossi, L. Navale, Y. Jiang, J. Aycock, M. Elias, D. Chang, J. Wiezorek and W. Y. Go, Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma, *New England Journal of Medicine* **377**, 2531 (December 2017).
3. J. Gust, R. Ponce, W. C. Liles, G. A. Garden and C. J. Turtle, Cytokines in CAR T Cell-Associated Neurotoxicity, *Frontiers in Immunology* **11**, p. 577027 (2020).
4. D. W. Lee, B. D. Santomasso, F. L. Locke, A. Ghobadi, C. J. Turtle, J. N. Brudno, M. V. Maus, J. H. Park, E. Mead, S. Pavletic, W. Y. Go, L. Eldjerou, R. A. Gardner, N. Frey, K. J. Curran, K. Peggs, M. Pasquini, J. F. DiPersio, M. R. van den Brink, K. V. Komanduri, S. A. Grupp

^eTARDIS is available open source at <https://github.com/LinguisticAnomalies/Coherence>

- and S. S. Neelapu, ASTCT Consensus Grading for Cytokine Release Syndrome and Neurologic Toxicity Associated with Immune Effector Cells, *Biology of Blood and Marrow Transplantation* **25**, 625 (April 2019).
5. B. D. Santomasso, J. H. Park, D. Salloum, I. Riviere, J. Flynn, E. Mead, E. Halton, X. Wang, B. Senechal, T. Purdon, J. R. Cross, H. Liu, B. Vachha, X. Chen, L. M. DeAngelis, D. Li, Y. Bernal, M. Gonen, H.-G. Wendel, M. Sadelain and R. J. Brentjens, Clinical and Biological Correlates of Neurotoxicity Associated with CAR T-cell Therapy in Patients with B-cell Acute Lymphoblastic Leukemia, *Cancer Discovery* **8**, 958 (August 2018).
 6. M. V. Maus, S. Alexander, M. R. Bishop, J. N. Brudno, C. Callahan, M. L. Davila, C. Diamonte, J. Dietrich, J. C. Fitzgerald, M. J. Frigault, T. J. Fry, J. L. Holter-Chakrabarty, K. V. Komanduri, D. W. Lee, F. L. Locke, S. L. Maude, P. L. McCarthy, E. Mead, S. S. Neelapu, T. G. Neilan, B. D. Santomasso, E. J. Shpall, D. T. Teachey, C. J. Turtle, T. Whitehead and S. A. Grupp, Society for Immunotherapy of Cancer (SITC) clinical practice guideline on immune effector cell-related adverse events, *Journal for ImmunoTherapy of Cancer* **8**, p. e001511 (December 2020).
 7. N. Möhn, V. Bonda, L. Grote-Levi, V. Panagiota, T. Fröhlich, C. Schultze-Florey, M. P. Wattjes, G. Beutel, M. Eder, S. David, S. Körner, G. Höglinger, M. Stangel, A. Ganser, C. Koenecke and T. Skripuletz, Neurological management and work-up of neurotoxicity associated with CAR T cell therapy, *Neurological Research and Practice* **4**, p. 1 (December 2022).
 8. P. Strati, S. Ahmed, F. Furqan, L. E. Fayad, H. J. Lee, S. P. Iyer, R. Nair, L. J. Nastoupil, S. Parmar, M. A. Rodriguez, F. Samaniego, R. E. Steiner, M. Wang, C. C. Pinnix, S. B. Horowitz, L. Feng, R. Sun, C. M. Claussen, M. C. Hawkins, N. A. Johnson, P. Singh, H. Mistry, S. Johncy, S. Adkins, P. Kebriaei, E. J. Shpall, M. R. Green, C. R. Flowers, J. Westin and S. S. Neelapu, Prognostic impact of corticosteroids on efficacy of chimeric antigen receptor T-cell therapy in large B-cell lymphoma, *Blood* **137**, 3272 (June 2021).
 9. M. M. Herr, G. L. Chen, M. Ross, H. Jacobson, R. McKenzie, L. Markel, S. R. Balderman, C. M. Ho, T. Hahn and P. L. McCarthy, Identification of Neurotoxicity after Chimeric Antigen Receptor (CAR) T Cell Infusion without Deterioration in the Immune Effector Cell-Associated Encephalopathy (ICE) Score, *Biology of Blood and Marrow Transplantation* **26**, e271 (November 2020).
 10. J. Gust, K. A. Hay, L.-A. Hanafi, D. Li, D. Myerson, L. F. Gonzalez-Cuyar, C. Yeung, W. C. Liles, M. Wurfel, J. A. Lopez, J. Chen, D. Chung, S. Harju-Baker, T. Özpölat, K. R. Fink, S. R. Riddell, D. G. Maloney and C. J. Turtle, Endothelial Activation and Blood-Brain Barrier Disruption in Neurotoxicity after Adoptive Immunotherapy with CD19 CAR-T Cells, *Cancer Discovery* **7**, 1404 (December 2017).
 11. J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li and F. Wei, Speech5: Unified-modal encoder-decoder pre-training for spoken language processing (2022).
 12. K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, J. Lane, M. Behrooz, W. Ngan, S. Poff, N. Goyal, A. Szlam, Y.-L. Boureau, M. Kambadur and J. Weston, Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage (2022).
 13. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez and I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena (2023).
 14. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra,

- I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov and T. Scialom, Llama 2: Open foundation and fine-tuned chat models (2023).
15. B. E. Gavett, A. S. Gurnani, J. L. Saurman, K. R. Chapman, E. G. Steinberg, B. M. Martin, C. E. Chaisson, J. Mez, Y. Tripodis and R. A. Stern, Practice effects on story memory and list learning tests in the neuropsychological assessment of older adults, *PLoS ONE* **11** (2016).
 16. M. Wilson and I. Division, Mrc psycholinguistic database: Machine usable dictionary, version 2.00., *Behav Res Methods* **20** (06 1997).
 17. M. D. Lezak and M. D. Lezak (eds.), *Neuropsychological assessment*, 4th ed edn. (Oxford University Press, Oxford ; New York, 2004).
 18. J. D. Herrera-García, I. Rego-García, V. Guillén-Martínez, M. Carrasco-García, C. Valderrama-Martín, R. Vílchez-Carrillo, S. López-Alcalde and C. Carnero-Pardo, Discriminative validity of an abbreviated Semantic Verbal Fluency Test, *Dementia & Neuropsychologia* **13**, 203 (June 2019).
 19. E. S. Gromisch, V. Zemon, R. H. Benedict, N. D. Chiaravalloti, J. DeLuca, M. A. Picone, S. Kim and F. W. Foley, Using a highly abbreviated California Verbal Learning Test-II to detect verbal memory deficits, *Multiple Sclerosis Journal* **19**, 498 (April 2013).
 20. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, Robust speech recognition via large-scale weak supervision (2022).
 21. S. Pakhomov, D. Chacon, M. Wicklund and J. Gundel, Computerized assessment of syntactic complexity in alzheimer’s disease: a case study of iris murdoch’s writing, *Behavior research methods* **43**, p. 136—144 (March 2011).
 22. T. Cohen and S. Pakhomov, A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the alzheimer’s type, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, eds. D. Jurafsky, J. Chai, N. Schluter and J. R. Tetreault (Association for Computational Linguistics, 2020).
 23. S. D. Canning, L. Leach, D. Stuss, L. Ngo and S. E. Black, Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia, *Neurology* **62**, 556 (February 2004).
 24. T. A. C. Thompson, P. H. Wilson, P. J. Snyder, R. H. Pietrzak, D. Darby, P. Maruff and H. Buschke, Sensitivity and Test-Retest Reliability of the International Shopping List Test in Assessing Verbal Learning and Memory in Mild Alzheimer’s Disease, *Archives of Clinical Neuropsychology* **26**, 412 (August 2011).
 25. S. Marino, S. Pakhomov, S. Han, K. Anderson, M. Ding, L. Eberly, D. Loring, C. Hawkins-Taylor, J. Rarick, I. Leppik, J. Cibula and A. Birnbaum, The effect of topiramate plasma concentration on linguistic behavior, verbal recall and working memory, *Epilepsy & Behavior* **24**, 365 (July 2012).
 26. S. V. Pakhomov, S. E. Marino, S. Banks and C. Bernick, Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency, *Speech Communication* **75**, 14 (December 2015).
 27. S. V. S. Pakhomov, W. Teeple, A. M. Mills and M. Kotlyar, Use of an automated mobile application to assess effects of nicotine withdrawal on verbal fluency: A pilot study., *Experimental and Clinical Psychopharmacology* **24**, 341 (October 2016).
 28. S. V. S. Pakhomov, L. E. Eberly and D. S. Knopman, Recurrent perseverations on semantic verbal fluency tasks as an early marker of cognitive impairment, *Journal of Clinical and Experimental Neuropsychology* **40**, 832 (September 2018).
 29. L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau and E. Coiera, Conversational agents in healthcare: a systematic review,

- Journal of the American Medical Informatics Association* **25**, 1248 (07 2018).
30. C. H. van Dyck, C. J. Swanson, P. Aisen, R. J. Bateman, C. Chen, M. Gee, M. Kanekiyo, D. Li, L. Reyderman, S. Cohen, L. Froelich, S. Katayama, M. Sabbagh, B. Vellas, D. Watson, S. Dhadda, M. Irizarry, L. D. Kramer and T. Iwatsubo, Lecanemab in early alzheimer's disease, *New England Journal of Medicine* **388**, 9 (2023), PMID: 36449413.
 31. B. Naughton, S. Saltzman, F. Ramadan, N. Chadha, R. Priore and J. Mylotte, A multifactorial intervention to reduce prevalence of delirium and shorten hospital length of stay, *Journal of the American Geriatrics Society* **53**, 18 (02 2005).
 32. P. Schramowski, C. Turan-Schwiewager, N. Andersen, C. Rothkopf and K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do, *Nature Machine Intelligence* **4**, 258 (03 2022).
 33. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Sementurs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* (July 2023).
 34. W. Xu, W. Wang, J. Portanova, A. Chander, A. Campbell, S. Pakhomov, D. Ben-Zeev and T. Cohen, Fully automated detection of formal thought disorder with time-series augmented representations for detection of incoherent speech (TARDIS), *J. Biomed. Informatics* **126**, p. 103998 (2022).