Leveraging Generative Artificial Intelligence to Develop Plain Language Summaries of Inpatient Progress Notes: A Quality Improvement Project

CN IRB Protocol Number: STUDY00001497

Principal Investigator: Dr. Kristen Elaine Johnson

Protocol Version Number: 1

4/21/2025

Table of Contents

LIST OF ABBREVIATIONS

PROTOCOL SUMMARY

Section 1 KEY ROLES

Section 2 INTRODUCTION, BACKGROUND INFORMATION, AND SCIENTIFIC RATIONALE
- 2.1 Background Information and Relevant Literature
- 2.2 Scientific Rationale
- 2.3 Potential Risks
- 2.4 Potential Benefits

Section 3 OBJECTIVES AND ENDPOINTS
- 3.1 Primary Objective(s)
- 3.2 Secondary Objective(s)
- 3.3 Primary Outcome Measure(s)
- 3.4 Secondary Outcome Measure(s)

Section 4 STUDY DESIGN

Section 5 STUDY ENROLLMENT AND WITHDRAWAL
- 5.1 Study Population, Recruitment, Retention
- 5.2 Inclusion Criteria
- 5.3 Exclusion Criteria
- 5.4 Vulnerable Subjects
- 5.5 Recruitment

Section 6 STUDY PROCEDURES
- 6.1 Informed Consent/Assent and HIPAA Authorization
- 6.2 Screening Process
- 6.3 Study Interventions and Follow-Up
- 6.4 Description of Study Procedures/Evaluations
- 6.5 Study Team Training and Intervention Reliability
- 6.6 Concomitant Interventions and Procedures

Section 7 STATISTICAL CONSIDERATIONS AND ANALYSIS
- 7.1 Statistical and Analytical Plans (SAPs)
- 7.2 Statistical Hypotheses

7.3 Analysis Datasets
7.4 Description of Statistical Methods
7.5 Sample Size
7.6 Measures of Minimize Bias

Section 8 DATA QUALITY AND OVERSIGHT
8.1 Study Team Quality Assurance and Quality Control
8.2 Data Safety and Monitoring Plan

Section 9 ETHICAL CONSIDERATIONS
9.1 Ethical Standard
9.2 Institutional Review Board (IRB)
9.3 Maintaining Subject Privacy
9.4 Maintaining Study Data Confidentiality
9.5 Study Support and Conflicts of Interest

Section 10 DATA HANDLING AND RECORD KEEPING
10.1 Data Management Responsibilities
10.2 Data Capture Methods
10.3 Study Record Retention Policy

Section 11 PUBLICATION POLICY

Section 12 REFERENCES

SUPPLEMENTS/APPENDICES

A: HIPAA Waiver
B: Overview for Residents
C: Evaluation Form
D: REDCap forms

List of Abbreviations

AI	**Artificial Intelligence**
API	**Application Programming Interface**
CNH	**Children's National Hospital**
CITI	**Collaborative Institutional Training Initiative**
CTSI	**Clinical and Translational Science Institute**
DSMB	**Data Safety Monitoring Board**
EMR	**Electronic Medical Record**
FKGL	**Flesch-Kincaid Grade Level**
HIPAA	**Health Insurance Portability and Accountability Act**
ICU	**Intensive Care Unit**
IRB	**Institutional Review Board**
LAR	**Legally Authorized Representative**
LLM	**Large Language Model**
MRN	**Medical Record Numbers**
OHRP	**Office for Human Research Protections**
PI	**Principal Investigator**
PHI	**Protected Health Information**
QI	**Quality Improvement**

Protocol Summary

Title:	Leveraging Generative Artificial Intelligence to Develop Plain Language Summaries of Inpatient Progress Notes: A Quality Improvement Project
Brief Summary:	We plan to use a Children's National's internal HIPAA-compliant large language model (LLM) to translate patients' progress notes into plain-language care summaries and evaluate them based on accuracy, readability, completeness, and tone.
Study Population:	Participants will be residents and attendings at Children's National Hospital who are reading and evaluating LLM-generated plain language summaries of progress notes for patients aged 0 - 22 years who are admitted to acute care at Children's National Hospital
Study Site(s):	Children's National Hospital
Number of Participants:	100 participants
Accrual Ceiling:	600 Participants
Study Duration:	May 15th 2025 - November 15th 2025
Subject Duration:	The patients and their families will not be active participants. Residents and attendings will spend 10-15 minutes reading and evaluating the care summaries.

Objective(s):	Our primary objective is to measure the accuracy of AI-generated plain language summaries of daily inpatient progress notes. Secondary objectives include measuring completeness, tone, and readability of these summaries.
Methodology:	Single-arm proof of concept quality improvement (QI) study
Outcome Measures:	Our outcome measures include scores for readability, accuracy, completeness, and tone.
Study Intervention/Procedures:	AI-generated care summaries will be created using Microsoft's Azure OpenAI Service, deployed within Children's National's Azure tenant. The resident who wrote the progress note and whenever possible an attending on the patient care team who signed the progress note will complete a survey to evaluate the care summary for its accuracy, completeness, and tone. We will also calculate readability using the Flesch-Kincaid Grade Level score.
Statistical Analysis:	We will analyze the metrics used to evaluate our AI-generated care summaries using descriptive statistics. Cohen's kappa statistic will be used to measure inter-rater agreement of accuracy, completeness, and tone between resident and attending assessments whenever possible.

Section 1: Key Roles

Principal Investigator (PI) /Oversight Monitor
Dr. Kristen Elaine Johnson, MD, MSC, Pediatric Hospitalist
Children's National Hospital
111 Michigan Ave, Washington, DC 22030
(571) 235-6524
KEJOHNSON@childrensnational.org

Primary Research Assistant
Miles Kaufman, B.S., Research Assistant
George Washington School of Medicine and Health Sciences
111 Michigan Ave, Washington, DC 22030
(312) 953-3125
miles.kaufman@gwu.edu

Study Team

- The study team will be responsible for screening patient charts, creating care summaries, recruiting residents, and data analysis.
- The study team consists of Dr. Johnson and Miles Kaufman
- Dr. Nisha Tamaskar will participate in data analysis.

Section 2: Introduction, Background Information, and Scientific Rationale

2.1 Background Information and Relevant Literature

Since the release of OpenAI's ChatGPT, almost every field has been exploring the potential of generative artificial intelligence (AI). Healthcare is no exception. Despite its promise, there are still significant gaps in research regarding its accuracy and benefits in clinical settings. This project aims to explore how generative AI can improve patient and family communication in pediatric inpatient settings through the use of a Health Insurance Portability and Accountability Act (HIPAA) compliant large language model (LLM).

Many hospitals, including Children's National, adopt a family-centered approach to morning rounds, actively involving patients and their families to encourage shared decision-making and ensure clarity regarding the care plan. This approach has been found to enhance communication and partnership between medical teams and families (Fernandes et al., 2021). However, families have reported that morning rounds can feel overwhelming due to the pace, number of people present, and complex medical jargon used (Fernandes et al., 2021). Also, whether intentional or not, it is common for physicians – especially trainees – to omit information during rounds (Artis et al., 2019), which could contribute to further confusion.

Research suggests that supplementing verbal communication with written care summaries can help to improve communication and satisfaction while reducing anxiety and depression for family members (Greenberg et al., 2022). Our proposed solution seeks to reduce stress and confusion for families by using an LLM to generate summaries of patient progress notes in clear, plain language.

This study will serve as a proof of concept, evaluating both the feasibility, accuracy, and utility of AI-generated care summaries. Prior studies have indicated that LLMs like ChatGPT are capable of generating medically accurate responses, but that they have the potential to "hallucinate", i.e. generate output that is inconsistent with the inputted information (Villarreal-Espinosa et al., 2024). Our primary question regarding feasibility is whether or not our LLM can reliably produce medically accurate output. Additionally, patient privacy concerns have limited real-world applications, with only 5% of studies using real patient data (Bedi et al., 2025), leading to a significant gap in research regarding AI's role in the care of real patients. Since Children's National's brand-new internal LLM is

both closed-loop and HIPAA-compliant, we will be able to use real patient data to determine if this LLM can be used to generate high-quality patient care summaries.

2.2 Scientific Rationale

We hypothesize that Children's National's internal HIPAA-compliant LLM will be able to accurately and effectively translate patients' progress notes into plain-language care summaries. This study is a pilot proof of concept, focusing on evaluating potential risks such as misinformation, omissions, or inappropriate tone in AI-generated summaries. Since no patient-facing interventions will occur, there is no indication for a traditional control or placebo group.

This study will focus on pediatric patients due to the unique involvement of families in pediatric care. Unlike in adult medicine, parents or guardians are often key participants in decision-making, making clear and accessible communication of medical information particularly important. Additionally, the use of AI in pediatric healthcare remains understudied (Bedi et al., 2025), and this project will help fill a critical gap in understanding how AI-generated summaries can enhance communication between providers and families.

Furthermore, Children's National has already invested time and effort into the LLM that we will be working with. This LLM must be studied to establish protocol and quality standards before Children's National begins to integrate this tool into clinical workflows. We hope to add valuable insight to its potential usefulness both in patient communication and resident education.

2.3 Potential Risks

Immediate risk

The main immediate risk is the potential for AI-generated content to influence healthcare providers in a way that negatively impacts patient care. Because the LLM is performing a summarization task rather than generating novel content or assisting in clinical decision-making, we believe there is low risk for this AI-generated content to negatively impact patient care.

However, to further safeguard against these AI-generated summaries influencing patient care, all clinicians involved will be explicitly reminded that the AI-generated content is not a validated medical resource and must not be used for clinical decision-making. This disclaimer will be communicated verbally before clinicians review the summaries and reinforced through a written disclaimer

included at the top of each AI-generated summary and the evaluation forms provided to clinicians.

Another potential risk is that residents may feel at risk for damage to professional reputation during the evaluation process. To mitigate this, we will emphasize that the study is a learning opportunity—encouraging residents to reflect on their notes and engage with emerging healthcare technology. We will also clarify to both residents and attendings that the quality of an AI-generated summary does not necessarily reflect the quality of the resident's original documentation. All evaluations of AI-generated summaries will remain confidential. The surveys of residents will be labeled with study IDs accessible only to the research team, and resident and attending names will not appear in the study data. We will make clear that participation is voluntary and will have no impact on resident evaluations.

Long-term risk

Breach of confidentiality is another risk to this study and for this reason, maintaining patient confidentiality is a critical consideration. While any use of patient data carries inherent risk, we will take all necessary precautions to ensure data security. Only the minimal amount of protected health information (PHI) required for the study will be used, and all data will be handled in accordance with HIPAA regulations and institutional security protocols.

Justification for potential risks

In addition to the justifications given above, AI's role in healthcare is expanding rapidly, and understanding how to integrate this technology safely and effectively is critical. One of the strengths of this study is the use of a HIPAA-compliant LLM, allowing for the analysis of real-world clinical data in a secure environment. Unlike many previous studies that relied on synthetic data, this approach will provide a more realistic evaluation of AI's impact in clinical settings. By studying AI's impact in a controlled, HIPAA-compliant environment, we aim to contribute valuable knowledge that will guide the responsible implementation of AI in patient communication.

2.4 Potential Benefits

There is no direct benefit to the patients whose progress notes are used in this pilot study. There is a societal benefit in that future patients may benefit from this research on how AI can be used to augment doctor-patient communication.

There is no direct benefit to the residents and attendings who participate in this study.

Section 3: Objectives and Endpoints

3.1 Primary Objective(s)

Our primary objective is to evaluate the overall accuracy of LLM-generated plain language summaries of daily inpatient progress notes using Children's National's newly developed HIPAA-compliant LLM.

3.2 Secondary Objective(s)

Our secondary objectives are to evaluate the completeness, tone, and readability of these LLM-generated plain language summaries.

3.3 Primary Outcome Measure(s)

1) **Accuracy**: Both the resident who wrote the progress note and when possible an attending physician will evaluate the accuracy of the medical content. Each inaccuracy will be categorized as an omission, an incorrect statement, or a hallucination which we define as extra information that is false or misleading (Kernberg, et. al., 2024). We will also ask the clinicians to evaluate the inaccuracies as either mild, moderate or severe. A mild inaccuracy is an inaccuracy that is unlikely to cause a serious misunderstanding, and a severe inaccuracy is an inaccuracy that could result in significant misunderstanding or harm.

3.4 Secondary Outcome Measure(s)

1) **Readability**: Assessed using the Flesch-Kincaid Grade Level (FKGL) score, which measures sentence length and word complexity.
2) **Completeness**: Both the same resident and when possible the attending on the medical service to whom the patient is admitted will evaluate the AI-generated summary, categorizing each key point as fully included (accurately captured without omission) partially included (some details missing or modified) or absent (omitted from the summary). The key points that we will ask the LLM to summarize are the patients' symptoms, medications, physical exam results, vitals, labs, imaging, clinical assessment, care plan, and consult orders.
3) **Tone**: The resident and, when possible, an attending will give the AI-generated care summary a score on a five point likert scale based on how much they agree with the following questions:
    a) The care summary reflects the patient's status appropriately.
    b) The care summary maintains a professional and appropriate tone.
    c) The information is presented in a way that the patient and/or their family could easily understand.
    d) I would feel comfortable allowing the patient and/or their family to read this document.
    e) Having AI-generated care summaries like this one would help me manage my workload more effectively

4) **Categorization of errors made by the LLM summarization progress:** Post-data collection, study team members Dr. Johnson and Dr. Tamaskar will review the original progress notes, LLM-generated summaries, and the attending/resident provider evaluations to categorize the types of errors identified in the summarization. Dr. Johnson and Dr. Tamaskar will review these errors separately in order to make independent judgments about the types of error identified. Identifying the categories of error will allow us to finetune the LLM output and improve the AI generated care summaries.

   We will include space for free-response feedback after each question about tone for the evaluation committee to elaborate on their assessment. This will be evaluated in order to help us finetune the LLM output and improve these AI generated care summaries.

Section 4: Study Design

This study is a single-center, single-arm proof of concept quality improvement (QI) study designed to evaluate the feasibility and accuracy of using a HIPAA-compliant LLM in generating plain-language care summaries from patient progress notes. The study will be conducted at Children's National Hospital and will focus on pediatric inpatient care.

Section 5: Study Enrollment and Withdrawal

5.1 Study Population, Recruitment, and Retention

- The target sample size is 100 evaluated AI care summaries. This target sample size was chosen based on sample size calculations with a goal of having 95% confidence that the true hallucination rate is less than 3% if zero hallucinations are observed, based on the Clopper-Pearson exact method.
- We plan to prescreen 600 patient charts because we anticipate that 500 charts will meet our inclusion criteria, and when we approach residents to ask if they will evaluate an AI-generated care summary, we anticipate that around 20% will agree.
- Recruitment will stop if we have prescreened 600 charts and successfully evaluated 100 care summaries, or if we reach 100 evaluated care summaries before pre screening 600 charts.
- All of the charts we review will be from patients currently admitted to Children's National Hospital's inpatient wards

5.2 Inclusion Criteria

- Participants:
    - Pediatric residents who authored or co-authored a daily inpatient progress note for a patient admitted to acute care at Children's National Hospital during the study inclusion period, or
    - Attending physicians who signed a daily inpatient progress note for a patient admitted to acute care at Children's National Hospital during the study inclusion period.
- Progress notes
    - Progress notes will be considered for LLM summarization if they were written for a patient hospitalized on an acute care floor

5.3 Exclusion Criteria

- Progress notes for patients admitted to the intensive care unit (ICU).

5.4 Vulnerable Subjects

This study involves summarizing the information of charts of pediatric patients, who are considered a federally recognized vulnerable population. The inclusion of children is necessary for this research, as the study aims to evaluate AI-generated care summaries in the context of pediatric inpatient care.

Pediatric populations are historically understudied in medical research, and this gap extends to the integration of artificial intelligence in healthcare. Generative AI is increasingly being explored in clinical settings, yet only 1.2% of studies have looked at the use of AI in pediatrics (Bedi et al., 2025). Studying how AI-generated care summaries function in pediatrics is critical to ensuring these tools can be safely and effectively implemented to enhance communication and improve patient and family experiences. Without targeted research in pediatric settings, there is a risk that AI tools may be developed without appropriate safeguards or optimizations for younger patients and their families.

We aim to responsibly conduct this research while maintaining the highest standards of patient protection. Since this study does not involve direct patient interventions, we believe that any risk to patient care will be minimal.

Our resident and attending physician evaluators may also be considered a potentially vulnerable population, given their roles as employees of Children's National. Because this project evaluates AI-generated summaries based on these clinicians' progress notes, they are uniquely qualified to assess the summaries for accuracy, completeness, and tone. To protect our participants, all evaluation responses will be kept strictly confidential, and no identifying information will be shared outside the study team. Evaluation forms and study materials will not be used in any way for performance assessments or employment-related evaluations.

The goal of this study is to better understand the limitations and capabilities of Children's National's generative AI model—a tool that may someday be integrated into clinical workflows. It is therefore essential to capture clinician perspectives. We are committed to ensuring that all study procedures are designed to minimize risk and protect our clinicians' privacy and professional standing.

5.5 Recruitment

Subjects will be recruited from residents and attendings working on the Children's National's inpatient wards. A HIPAA waiver will be requested from the IRB to pre-screen medical records to identify potential progress notes for summarization.

All pre-screening will be conducted using Children's National's electronic medical record (EMR), which is maintained on password-protected computers on the Children's National Hospital network. Only authorized study team members will have access to the pre-screening materials. If the patient does not meet inclusion criteria, no PHI will be taken from the EMR.

Section 6: Study Procedures

6.1 Informed Consent/Assent and HIPAA Authorization

Since this is a quality improvement project assessing the effectiveness of Children's National's LLM, we are requesting a waiver of informed consent for both the patients whose progress notes will be used and the clinicians who will be evaluating the AI-generated summaries.

We believe a waiver of consent is justified given the large number of patient notes required to assess care summary quality. Patients will not interact with the AI-generated summaries or the study team.

If a progress note meets our inclusion criteria, we will begin by pulling basic demographic information about the patients. This will all be kept secure in RedCap. All care summaries, which may contain PHI, will be kept in a secure Children's National OneDrive folder, and only the evaluation metrics, which contain no PHI, will ever be exported for analysis. As such, the risk to patients is minimal, and seeking informed consent would be unnecessary and burdensome to both the study team and families.

We also believe a waiver of consent is justified for clinician participants. An information sheet will be provided explaining the purpose and procedures of the QI project. They will have the opportunity to ask questions before agreeing to participate and will be informed that their participation is entirely voluntary, that they may opt out at any time without consequence, and that study results will not be used for performance evaluations. Furthermore, the duration of their participation will be limited to the 10-15 minutes it takes to evaluate the care summary.

The only personal information we will be recording will be names for administrative purposes in case we need to clarify parts of their evaluation or rectify any errors. Clinicians' names will be kept in RedCap and marked as identifiers so that they are kept securely not extracted for data analysis. Our RedCap project and secure Children's National OneDrive folders will only be accessible to study team members. No personal information about clinicians will be used in our analysis. Therefore we believe this quality improvement study qualifies for exemption from informed consent.

6.2 Screening Process

We have requested a HIPAA waiver from the institutional review board (IRB) to pre-screen medical records to identify potential study participants.

6.3 Study Interventions and Follow-Up

This study includes no direct interventions or follow-ups with patients.

6.4 Description of Study Procedures/Evaluations

Once a patient progress note has been deemed eligible, we will take the following steps:

1) Generating the care summary
    a) The most recent progress note will be copied and entered into our application programming interface (API), which securely connects to Children's National's HIPAA-compliant generative AI platform.
    b) The API includes specific instructions for the LLM on how to summarize a clinical progress note into plain language.
    c) The summary will be generated by Microsoft's HIPAA-compliant service, which grants Children's National access to OpenAI's suite of generative AI models. Importantly, neither the original note nor the resulting care summary is stored by the model or platform at any time.
2) Readability Scoring and Data Entry
    a) We have written a python script that will Our API will also calculate a readability score using the py-readability-metrics 1.4.5 Python package. This score will be recorded in REDCap, and the care summary text itself will not be stored by the Python script.

b) The AI-generated care summary will be copy and pasted into a word document and then uploaded as a PDF to a secure Children's National OneDrive folder, these files will be labeled with studyID codes that are linked to the patient's MRN in REDCap. Access to this OneDrive folder and REDCap will be restricted to the study team only.
c) The original progress note will not be retained or stored.

3) Clinician Evaluation Process
   a) Two printed copies of the care summary will be prepared—one for the resident who authored the note, and one for the attending physician (when available).
   b) A member of the study team will approach the resident (and attending, when possible) to invite them to evaluate the summary. Each clinician will be provided with an information sheet outlining the study purpose, procedures, and their rights as participants. Participation is entirely voluntary and confidential.
   c) If a clinician declines, their corresponding AI-generated summary will be immediately shredded.
   d) If a clinician agrees to participate, they will complete a paper-based evaluation form and may mark any errors or issues directly on the printed care summary.
4) Evaluation Data Handling and Storage
   a) Study team members will enter evaluation results into REDCap. Both the completed evaluation form and the marked-up care summary will be scanned and uploaded into a secure Children's National OneDrive folder for auditing purposes.
   b) Access to evaluation forms will be restricted and any PHI entered into REDCap will be marked as such . However, the numerical scores and responses used in data analysis will be de-identified.
   c) Clinician names will be stored in REDCap solely for administrative purposes and will be flagged as identifiers. Each clinician will also be assigned a unique study ID, linked to their name in a password-protected codebook stored on a secure Children's National drive.
   d) All printed summaries and evaluation forms will be securely shredded to minimize PHI exposure. While we aim to scan, upload, and shred all materials the same day, all paper documents will be destroyed no later than 10 days from creation.
5) Collecting Patient Demographic Data
   a) Once we have an evaluation form for an AI care summary we will also enter basic demographic information about the patient into a RedCap form. This information will include sex, race, ethnicity, age, payer type, date of admission and date of progress note, and primary reason for admission as indicated by the principal diagnosis of the problem list.

b) These data will help ensure diversity in the sample and allow for assessment of potential model bias. All identifiable fields will be appropriately marked and stored securely.

6) Data Security and Retention
   a) All REDCap data, files in OneDrive and codebooks will be maintained either in REDCap or on secure institutional drives. These files will be permanently deleted three years after study completion.
   b) Any future use of the de-identified data will require separate IRB approval.

6.5 Study Team Training and Intervention Reliability

Miles Kaufman, a first-year medical student on this project, will be responsible for obtaining consent from patients and families. She has completed the Collaborative Institutional Training Initiative (CITI) Good Clinical Practice and Social & Behavioral Research training modules and has a valid HIPAA certification from George Washington University. She will be under the supervision of Dr. Kristen Johnson, who will also ensure compliance with HIPAA regulations and best research practices.

Dr. Nisha Tamaskar, a pediatric hospitalist attending at Children's National, will collaborate on error analysis with Dr. Johnson. Dr. Tamaskar has completed the Collaborative Institutional Training Initiative (CITI) Good Clinical Practice and Social & Behavioral Research training modules and has a valid HIPAA certification.

To ensure our LLM reliably produces high-quality care summaries, the study team will first evaluate the model using 20 fictitious patient progress notes that have been generated by ChatGPT using prompts designed by the research team. We will only move to real patient data once the summaries meet the following benchmarks:

1) Average reading level between 4th and 8th grade
2) Inclusion of at least 5 out of 6 key sections on average
3) No severe inaccuracies, fewer than 5 moderate inaccuracies total, and fewer than 10 mild inaccuracies total across all summaries
4) The study team evaluators all "agree" or "strongly agree" that the care summaries are written in a professional and appropriate tone

Once the API meets these criteria, the code will remain unchanged for the duration of the study. We will re-evaluate the model after generating summaries for 20 and 40 real patients to ensure it continues to meet these standards, and if it does not, then we will finetune our API and restart our protocol.

6.6 Concomitant Interventions and Procedures

Care should be provided as normal for all of our study subjects and residents should be evaluated by their supervising attendings normally. We will make this clear to any and all participants in our study. We will emphasize both verbally and written to all involved clinicians that the AI models we are using are currently untested and that clinical care should not be based on any information from the AI-generated care summaries.

Section 7: Statistical Considerations and Analysis

7.1 Statistical and Analytical Plans

As a proof of concept quality improvement project, this study will not have a formal statistical analysis plan.

7.2 Statistical Hypotheses

Our primary hypothesis is that AI-generated care summaries will be accurate, written for a lay audience, include all major points from the progress note, and use an appropriate tone. For the purposes of sample size calculations, we focus on the inference we can draw about the true hallucination rate if no hallucinations are detected in the final product.

Let:

- p = true hallucination rate of the study LLM
- p_0 = hallucination rate of 3% (i.e., 3% of LLM-generated summaries contain any hallucination)

We are testing:

$H_0: p > p_0$

$H_A: p \leq p_0$ (the hallucination rate is acceptably low, i.e. less than 3%)

We use the Clopper-Pearson exact binomial confidence interval to assess this. If the upper bound of the 95% confidence interval for the observed proportion of hallucinations is less than or equal to p_0, we reject the null hypothesis and conclude with 95% confidence that the hallucination rate is less than 3%.

7.3 Analysis Datasets

We will conduct a per-protocol analysis of all AI-generated care summaries. Each summary will be evaluated by both the resident who authored the patient's

progress note and whenever possible an attending on the patient care team who cosigned the progress note.

In REDCap, we will store both PHI and data intended for analysis. The following elements will be flagged as PHI and excluded from all data analysis in this phase of the study:

-
- Identifiable demographic information such as patients or clinician names, medical record numbers (MRNs), birth dates, primary reason for admission, and admission dates.
- Any long form feedback or written short responses on the evaluation forms will be flagged until they are thoroughly de-identified.

Data we will analyze includes:

- Non-identifiable demographic information such as sex, race, age, length of admission, and payer type.
- Quantitative evaluation metrics such as readability scores, completeness ratings, counts and types of inaccuracies, and Likert scale responses assessing tone
- We will conduct a thematic analysis of the types of errors identified by the attending and resident reviewers of the plain language summaries. Dr. Johnson and Dr. Tamaskar will independently review each care summary, the original progress note, and the participant evaluation, and conduct a thematic analysis on the errors identified by participants. After Drs. Johnson and Tamaskar independently review and categorize the errors, they will meet to discuss the categories of errors they assessed and identify ways that the LLM could be fine tuned to further reduce error.
- We will conduct a thematic analysis of the de-identified free-text responses.

7.4 Description of Statistical Methods

This is a single-arm proof-of-concept study. We will collect baseline demographic data to assess the representativeness of our sample and identify any potential gaps in diversity. After generating care summaries, we will evaluate their readability, completeness, accuracy and tone using the metrics outlined in section 3.3 and the free-response questions outlined in section 3.4.

All numerical data will be analyzed using descriptive statistics. Open-ended responses will undergo qualitative thematic analysis to extract key insights regarding content sensitivity, potential benefits for families, and areas for improvement. Cohen’s Kappa statistic will be used to assess inter-rater agreement between resident and attending assessments of a summary, whenever applicable.

We will conduct interim analyses after evaluating care summaries for 20 and then 40 patients to ensure that the summaries are meeting the benchmarks defined in section 6.5. Given the exploratory nature of this study, it is essential that we have this opportunity to finetune our API or train the LLM if necessary. Once data collection is complete, the study team will conduct the final statistical analysis.

7.5 Sample Size

Using the Clopper-Pearson Confidence interval function in R (Posit Team, 2023), we determine the minimum number of summaries required such that the upper bound of the 95% confidence interval for the observed proportion is ≤ 3%, given that 0 hallucinations are observed in the study population of summaries. The results of this function indicate that 100 summaries are needed to be able to have 95% confidence that the true rate of hallucinations is less than 3%, if 0 hallucinations are observed in the LLM-generated summaries.

7.6 Measures to Minimize Bias

As a single-arm proof-of-concept study, we will not implement blinding or randomization at this stage. To minimize bias, clinicians will independently evaluate the care summaries based on standardized instructions. For many care summaries we hope to have both a resident and an attending physician evaluate the summary in order to reduce bias and catch inconsistencies.

Section 8: Data Quality and Oversight

8.1 Study Team Quality Assurance and Quality Control

The research records including the codebooks, files in OneDrive and REDCap data will be accessible only to authorized study personnel listed on this IRB. The team will monitor REDCap to ensure data is being entered accurately. Access will be granted in accordance with institutional policies and regulatory requirements to ensure data integrity and confidentiality. Identified protocol deviations will be logged and reported promptly to Dr. Johnson. Any breach in confidentiality will be reported to the IRB. Per Children's National Hospital (CNH) requirements, all protocol deviations will be reported within seven (7) business days of discovery.

8.2 Data Safety and Monitoring Plan

We have planned interim analyses after 20 and 40 summaries are evaluated. This will ensure that our LLM is performing above our preset tolerance level for readability, completeness, accuracy, and tone (section 6.5). If our LLM is not performing at that level, we will alter our API, potentially retrain our LLM and restart our protocol once the issue has been resolved.

If major discrepancies arise between the resident and attending evaluations, the study team will reassess the data. Erroneous grader input will be excluded from the final analysis. Dr. Johnson and the study team will be responsible for ensuring data quality, implementing corrective actions, and ensuring compliance with data management procedures.

Section 9: Ethical Considerations

9.1 Ethical Standard

The study team will ensure that this study is conducted in full conformity with the Regulations for the Protection of Human Subjects of Research codified in 45 Part 46 of the Code of Federal Regulations, Children's National Policies and Procedures and Good Clinical Practices.

9.2 Institutional Review Board

The protocol, HIPAA waiver, recruitment materials, and all participant materials will be submitted to the Children's National IRB for review and approval. Approval of both the protocol and the evaluation form will be obtained before we begin generating care summaries using patient progress notes. Any change to the protocol, consent, recruitment materials, and participant information sheets or letters will require IRB approval before implementation and use. The IRB will determine whether previously consented participants need to be re-consented and whether consent of more than one parent is required for minors.

The IRB will be notified of study team updates via an amendment. Data Safety Monitoring Board (DSMB) Reports will be submitted at the time of the continuing review or with another applicable IRB transaction.

Other study events (e.g., protocol deviations, data monitoring reports) will be submitted per the Children's National IRB Reportable Events Module.

9.3 Maintaining Subject Privacy

The patient progress notes we use will not be saved. They will be imputed into the HIPAA compliant LLM that is managed by the hospital's AI and Advanced Analytics team. The output (progress note summaries) will be copy and pasted into a word document and saved as a PDF to a secure Children's National OneDrive folder. No participant data will be stored on the platform where the LLM is hosted.

All care summaries will be printed and evaluation forms will be filled out on paper. Once the evaluation form is uploaded into OneDrive, both the printed care summary and evaluation form will be shredded.

All study data including patient demographic data and manually entered evaluation data will be managed using HIPAA-compliant, password-protected software namely REDCap. All files including scans of paper evaluation forms and PDFs of the AI generated care summaries will be stored in a secure Children's National OneDrive. All PHI stored in REDCap will be marked as such and we will disable settings that would allow any team member to download or extract PHI. The codebook linking clinicians' names to their study IDs and linking patient MRNs to their studyIDs will be stored on a secure Children's National drive.

9.4 Maintaining Study Data Confidentiality

Participant confidentiality is strictly held in trust by the participating investigators and the study team. This confidentiality is extended to cover all clinical information relating to participants and clinician notes. Therefore, the study protocol, documentation, data, and all other information generated will be held in strict confidence.

No information regarding the study or its data will be disclosed to any unauthorized third party without prior written approval from the patients' legally authorized representative (LAR) and the IRB. Study data, including a codebook linking study IDs to participants' names will be accessible only to the study team, including Dr. Kristen Elaine Johnson, a pediatric hospitalist, and Miles Kaufman, a first-year medical student and research team member. This codebook will be kept in a secure Children's National OneDrive folder drive and permanently deleted three years after study completion.

The regulatory authorities (e.g., IRB, Office for Human Research Protections (OHRP)) may inspect all documents and records required to be maintained by the investigator. Children's National will permit access to such records.

The study data entry and study management systems used by research staff will be secured and password-protected. After the study has concluded, all research records will be securely stored in REDCap for three years at which point, all study databases will be archived at Children's National.

The research data will not contain participants' contact or identifying information. Instead, participants and their data will be identified by unique study identification numbers. While demographic information will be included in the research data, it will not be reported at the individual level but will be summarized using descriptive statistics to characterize the overall cohort.

9.5 Study Support and Conflicts of Interest

Miles Kaufman is receiving a $4000 grant provided by the Gill Fellowship – a stipend for summer research awarded to first-year medical students at George Washington School of Medicine and Health Sciences. REDCap® support is provided by The Clinical and Translational Science Institute (CTSI) at Children's National. Access to the LLM software we will be using will be provided by Children's National Hospital via their established relationship with Microsoft.

All key study personnel will follow the Human Research Protections Program Investigator, Study Staff, and Family Member Conflicts of Interest Policy.

Section 10: Data Handling and Record Keeping

10.1 Data Management Responsibilities

The study team, including Miles Kaufman and Dr. Kristen Elaine Johnson, will be responsible for pre-screening patients, recruiting residents and attendings, and overseeing the evaluation of AI-generated care summaries. Team members will have access to the REDCap database and the secure Children's National OneDrive folder containing care summaries, evaluation forms and the codebook linking study IDs to participants.

Responsibilities of the study team include:

- Entering patient demographic information
- Maintaining a secure codebook linking study IDs to clinician identifiers
- Uploading AI-generated care summaries on to OneDrive
- Entering readability scores calculated using the standardized Flesch-Kincaid Grade Level formula into REDCap
- Printing AI generated care summaries and recruiting clinicians to evaluate them. Providing evaluators with clear, standardized instructions to ensure accurate and consistent assessment of the AI-generated summaries
- Scanning and uploading evaluation forms and graded care summaries onto OneDrive
- Inputting evaluation metrics into a REDCap form
- Ensuring care summaries and evaluation forms are all shredded after the evaluation is complete
- Once data collection is complete, the study team will be responsible for analyzing de-identified data using R and Python's statistical software.

10.2 Data Capture Methods

Data collection will be conducted by the study team under the oversight of the PI, who is responsible for ensuring the accuracy, completeness, legibility, timeliness, and completeness of the data reported.

Children's National will maintain all relevant source data. Source data include all information and original records of clinical findings, observations, or other activities necessary for the reconstruction and evaluation of the trial. Examples of original source documentation include electronic medical records, laboratory reports, memoranda, subject diaries, patient progress notes, subject questionnaires, and recorded data from automated instruments.

AI-generated care summaries will be created using Microsoft's Azure OpenAI Service, deployed within Children's National's Azure tenant. This is a HIPAA-compliant and secure service that provides access to OpenAI's suite of Generative AI models. This service is a part of Children's National's Azure tenant, and is managed by the hospital's AI and Advanced Analytics team with whom our study team will be working closely. All API development, data input, and output handling will occur on password-protected, approved computers connected to the secure Children's National network. Moreover, using the OpenAI service requires an API key, which is similar to a password. The AI and Advanced Analytics team will manage this API key and use Microsoft's recommended approaches to facilitate the research team's access while preserving key encryption. No patient data will be stored long-term by the LLM or OpenAI service. All AI-generated care summaries we retain will be stored on a secure Children's National OneDrive.

All study data will be entered into REDCap®, a password-protected, secure, HIPAA-compliant, web-based electronic database with a built-in audit trail. (Harris et. al., 2009)

Only IRB approved research team members who have current HIPAA and CITI Good Clinical Practice and human subjects protection training will be authorized to extract data from source documents and enter it into REDCap®.

Data will be entered directly from the source documents into REDCap® or uploaded into the OneDrive folder within 10 days of collection.

10.3 Study Record Retention Policy

Study records will be maintained for 3 years after study closure as per HHS 45 CFR part 46.115b at which point digital files will be deleted. Paper documents will all be shredded after use, no later than 10 days after they are initially generated.

Section 11: Publication Policy

Dr. Kristen Johnson will be responsible for developing publication procedures and resolving authorship issues.

Section 12: References

Artis, Kathryn A et al. “Data Omission by Physician Trainees on ICU Rounds.” Critical care medicine vol. 47,3 (2019): 403-409. doi:10.1097/CCM.0000000000003557

Bedi, Suhana et al. “Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review.” JAMA vol. 333,4 (2025): 319-328. doi:10.1001/jama.2024.21700

Fernandes, Ashley K et al. “Pediatric Family-Centered Rounds and Humanism: A Systematic Review and Qualitative Meta-analysis.” Hospital pediatrics vol. 11,6 (2021): 636-649. doi:10.1542/hpeds.2020-000240

Greenberg, Jared A et al. “Daily Written Care Summaries for Families of Critically Ill Patients: A Randomized Controlled Trial.” Critical care medicine vol. 50,9 (2022): 1296-1305. doi:10.1097/CCM.0000000000005583

Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, Jose G. Conde, Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377-81.

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software,
PBC, Boston, MA. URL http://www.posit.co/.

Kernberg, A., Gold, J. A., & Mohan, V. (2024). Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. Journal of medical Internet research, 26, e54419.

Villarreal-Espinosa, Juan Bernardo et al. “Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery.” The Knee vol. 51 (2024): 84-92. doi:10.1016/j.knee.2024.08.014