Abstract

Introduction

Materials & Methods

Results

Conclusion

SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

*equal contribution

†equal contribution

[1]Department of Medicine, BIDMC

[2]Information Systems, BILH

[3]Halo Solutions, LLC

Summary

Adoption of a new electronic health record (EHR) affords a unique opportunity for improving quality related workflows. To fully leverage novel capabilities, such as population health reminders for colonoscopy recall, heterogeneous legacy data in the form of text notes and PDFs must be structured at scale. Here, we describe the use of state-of-the-art Large Language Models (LLMs) to perform population-wide colonoscopy recall inference and compare this to a patient-naïve default 10-year interval.

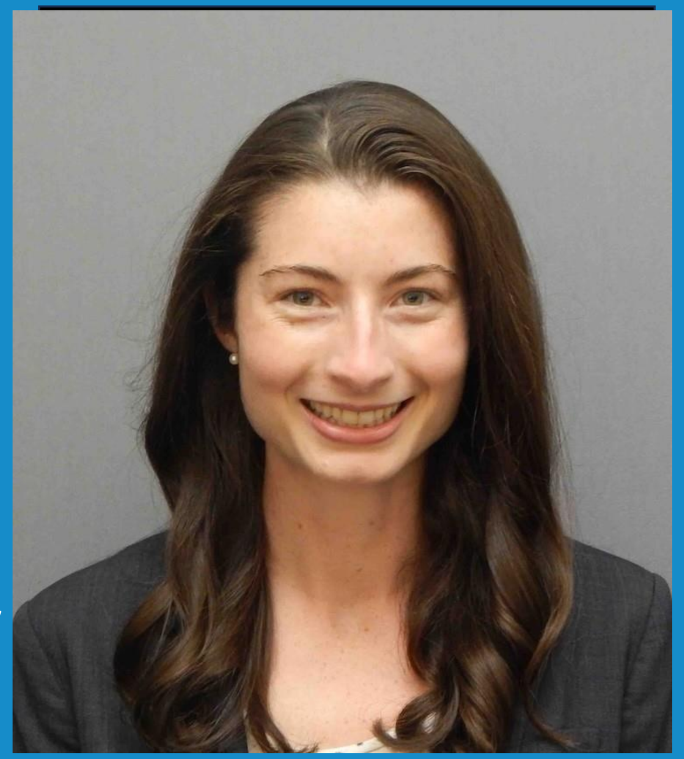
Meet the Authors

Aman Mohapatra

Aman is an Internal Medicine PGY1 at Beth Israel Deaconess Medica Center (BIDMC) interested in Gastroenterology. Current research interests include the intersection between Artificial Intelligence and clinical decision making, IBS, and Celiac Disease.

Rachel Porth

Rachel is an Internal Medicine PGY3 at Beth Israel Medical Center (BIDMC) who just matched into Gastroenterology at BIDMC. Current research interests include Inflammatory Bowel Disease, utilities of Artificial Intelligence, and quality improvement.



Beth Israel Lahey Health

HARVARD MEDICAL SCHOOL TEACHING HOSPITAL

Halo Solutions

SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

[*]equal contribution
[†]equal contribution
[1]Department of Medicine, BIDMC
[2]Information Systems, BILH
[3]Halo Solutions, LLC

Abstract

Introduction:
Transitioning electronic medical records (EMR) and transferring patient data affords a unique opportunity for improving quality related workflows. At our academic medical center, we transitioned to Epic (an EMR) with the inability to seemlessly transfer preventative health patient data from our old EMR – exposing a potential care-gap for high-risk disease patients. To fully leverage novel capabilities, such as population health reminders for colonoscopy recall, heterogeneous legacy data in the form of text notes and PDFs must be structured at scale. Here, we describe the use of state-of-the-art Large Language Models (LLMs) to perform population-wide colonoscopy recall inference and compare this to a patient-naïve default 10-year interval.

Methods:
Historical colorectal cancer (CRC) screening/surveillance data (reports, scheduling, letters) were compiled for our 700-bed tertiary academic medical center. An LLM-based recall workflow was developed iteratively involving Optical Character Recognition (Azure Document Intelligence) of procedure report PDFs, GPT-4 / GPT-4 Turbo (Azure Cloud endpoints) for recall inference from reports and follow-up letters (with preference given to recall from letters), and custom workflows for outlier detection and filtering (Halo Solutions).

Results:
118,061 patients with historical colonoscopy data from 2014-2024 were identified for recall inference. The mean age at last colonoscopy was 59 +/-13, consistent with other real-world cohorts (with an increasingly prominent peak at age 45 coinciding with the 2021 lowering of the recommended screening age by the USPSTF). Initial prompting of GPT-4 demonstrated excellent performance (100/100 concordant recalls compared to clinician review; Macro F1=1.0). Further testing on 20 potentially more challenging cases from bidirectional endoscopies surfaced one failure (misinterpretation of recall age as a recall interval) that necessitated additional prompt engineering as well as upgrading to GPT4-Turbo. After an additional round of validation and optimization, population-wide inference was performed. Post hoc quality control confirmed canonical patient recall trajectories (IBD/CRC family history patients, symptomatic patients with early diagnostic endoscopies, and traditional risk-based surveillance patients) and motivated additional filtering for outliers (e.g., intervals <6m or >10y). To promote usability, a comment noting the recall was AI-generating and citing the date of the report/note used was included, and continuous post-deployment feedback was collected. Relative to a naïve 10-year interval, our LLM-based workflow is estimated to prevent over 400 new cancers.

Discussion:
Use of an LLM-based workflow for colonoscopy recall inference proved to be feasible at population scale. Key lessons from implementation include the importance of stakeholder alignment, strategies for recall specific prompt engineering, methods for post-hoc quality control of LLM results at scale, and solutions to enhance the usability of inference results and enable continuous feedback.

SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

[*]equal contribution
[†]equal contribution
[1]Department of Medicine, BIDMC
[2]Information Systems, BILH
[3]Halo Solutions, LLC

Introduction

- Epic, a type of electronic medical record, has a tab called "the health maintenance tab" which stores preventative patient data such as next colonoscopy, mammogram, and vaccine information.
- The biggest problem with the health maintenance tab in Epic is its inability to transfer data from alternate EMRs and its need to be manually updated by a provider – leaving room for user errors and care gaps.
- The transition to a new EMR system leaves many hospitals with the difficulty to transfer patient data. Unfortunately, this could create a care gap for thousands of patients with preventative health issues.
- At our tertiary academic medical center, we transitioned to Epic in June 2024. At the time of the transition, there was not an existing solution to transfer health maintenance data from our previous electronic medical record.
- The Problem: a default approach of assigning a universal 10-year recall interval for all patients based on their last recorded colonoscopy date may lead to delays in screening for high risk patients.
- The Solution: We designed a system-wide workflow to structure colonoscopy recall guidance, enabling accurate carry over of historical screening and surveillance recommendations in the new medical record system
- Our workflow leveraged recent advances in LLMs to structure all historical data types and provide an adaptive, clinically informed recall interval based on a patient's last colonoscopy encounter.

Figure 1. LLM Colonoscopy Recall Workflow Diagram

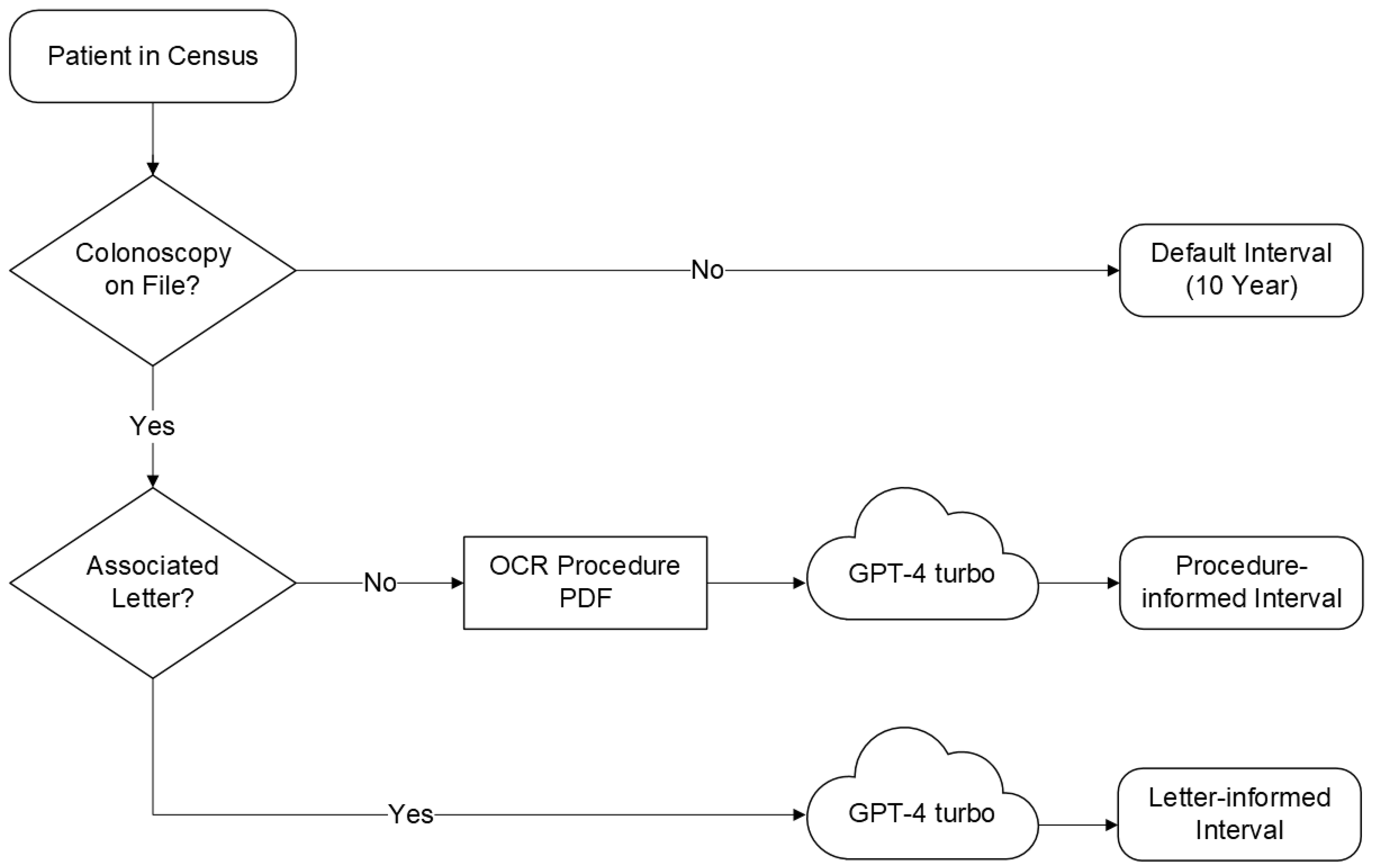


Fig 1. This diagram shows the schema used to implement an LLM driven recall application using patient letters and procedure reports for extraction of accurate colonoscopy recall.

Home

Miake-Lye et al. Transitioning from One Electronic Health Record to Another: A Systematic Review. J Gen Intern Med. 2023.

SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

[*]equal contribution
[†]equal contribution
[1]Department of Medicine, BIDMC
[2]Information Systems, BILH
[3]Halo Solutions, LLC

Materials and Methods

Stakeholders

- hospital leadership, information systems, gastroenterology, quality and safety, and the analytics service provider (Halo Solutions).

Optimization of LLM-Based Recall Inference

- Used patient colonoscopy letters and procedure reports.
- Tested GPT-4 with small sample and manually reviewed.
- Further refinement through challenging cases, with endoscopy reports included.
- Compared GPT-4 with GPT-4 Turbo for reduction of errors.

Scaling and Outlier Detection

- Ran GPT-4 Turbo workflow on 1000 patients, and performed group aggregate analysis to detect outliers.
- Manually examined outliers and adjusted GPT-4 Turbo recall parameters.
- If recall >10 years and patient is <45 – the data was filtered out.
- If recall was <6 months – the data was filtered out.

Real World Deployment and Continuous Feedback

- For transparency – included an automatically generated comment indicating the source and date of the recommendation.
- For patients 80 and above – added boilerplate text that will inform physicians to incorporate a shared decision-making process.
- Following deployment, continuous feedback was solicited from primary care physicians to gain insight into usability.

Home

SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

Results

Table 1. Patient Demographics and Recall Sources

Feature	N (% out of 118,061)
Age	
<40	6451 (5.5%)
40-49	8371 (7.5%)
50-59	25596 (22%)
60-70	35334 (30%)
70-79	28244 (24%)
≥80	13530 (11%)
Gender	
Male	55414 (47%)
Female	62646 (53%)
Race	
Caucasian	64484 (54%)
Asian	11062 (9.3%)
Black	9930 (8.4%)
Hispanic	5200 (4.4%)
Unknown/Other	27385 (23%)
Recall Sources	
Procedure Report	54660 (46%)
Procedure Report + Letter	63401 (54%)

Table 2. Performance of GPT-4 on Colonoscopy Recall Task

Category	Total Cases	Correctly Identified	Errors
Procedure Reports	100	100	0
Patient Letters	57	57	0
Complex Cases (EGD & Colonoscopy)	20	19	1

Purpose: Aggregate analysis of population-wide data using GPT-4 Turbo due to infeasibility of manual review at scale (Figure 2).

Identified Groups:

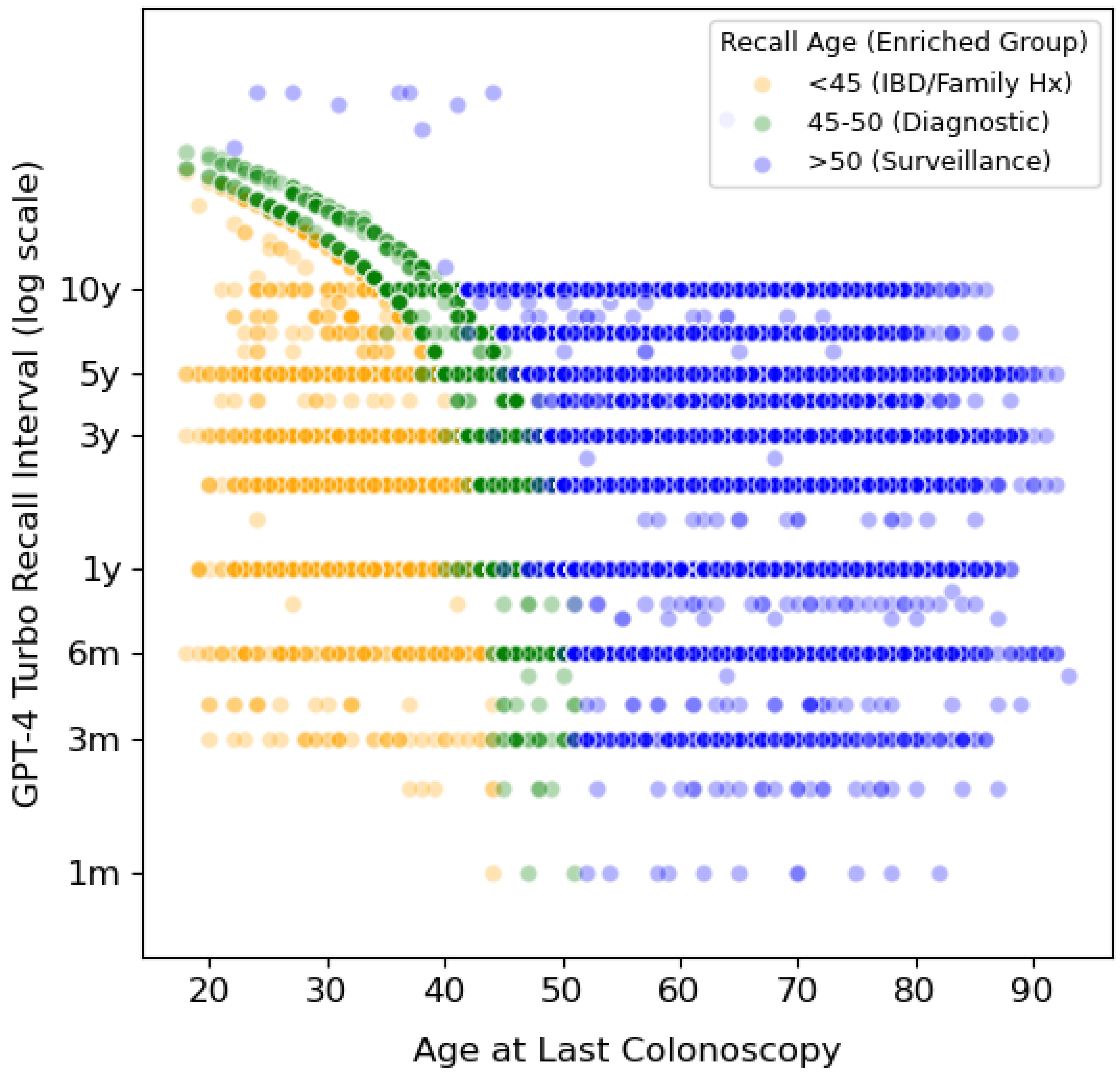
- **Under 45**: Patients with IBD or family history, typical routine screening group.
- **Ages 45–50**: Diagnostic follow-ups planned for age-appropriate screening cutoffs.
- **Over 50**: Standard risk-based surveillance group.

Outliers:

- **>10-Year Recalls**: Errors caused by misinterpreted recall intervals as calendar years (e.g., “2026 years”).
- **<6-Month Recalls**: Reflecting clinical follow-up rather than next colonoscopy.

Significance: By running the LLM on the entire dataset and stratifying the patients into three clinically relevant recall groups – outliers could be detected at a scale infeasible by manual review of LLM output (>100K patients).

Figure 2. Post Hoc Quality Control for Outlier Detection in Population-Wide LLM Recall Workflow (N = 111,380 Patients with Recalls)



SYSTEM WIDE IMPLEMENTATION OF A LARGE LANGUAGE MODEL WORKFLOW FOR COLONOSCOPY RECALL INFERENCE

Authors: Rachel Porth[1*], Aman Mohapatra[1*], Si Wong[2], Heather Hardway[3], Gail Piatkowski[2], John Shang[2], Maelys J. Amat[1], Sarah Flier[1], Adam Salsman[2], Ted Fitzgerald[2], Ayad Shammout[2], David Rubins[1,2], Amy Miller[1,2], Venkat Jegadeesan[2], Arvind Ravi[3†], Joseph D Feuerstein[1†]

[*]equal contribution
[†]equal contribution
[1]Department of Medicine, BIDMC
[2]Information Systems, BILH
[3]Halo Solutions, LLC

Conclusion

Problem Addressed: The transition to Epic at our academic center highlighted the potential large care gaps for high-risk patients due to the absence of a straightforward way to transfer colonoscopy recall data, leading us to design an LLM-based workflow to integrate historical colonoscopy data from the old EMR to the new EMR.

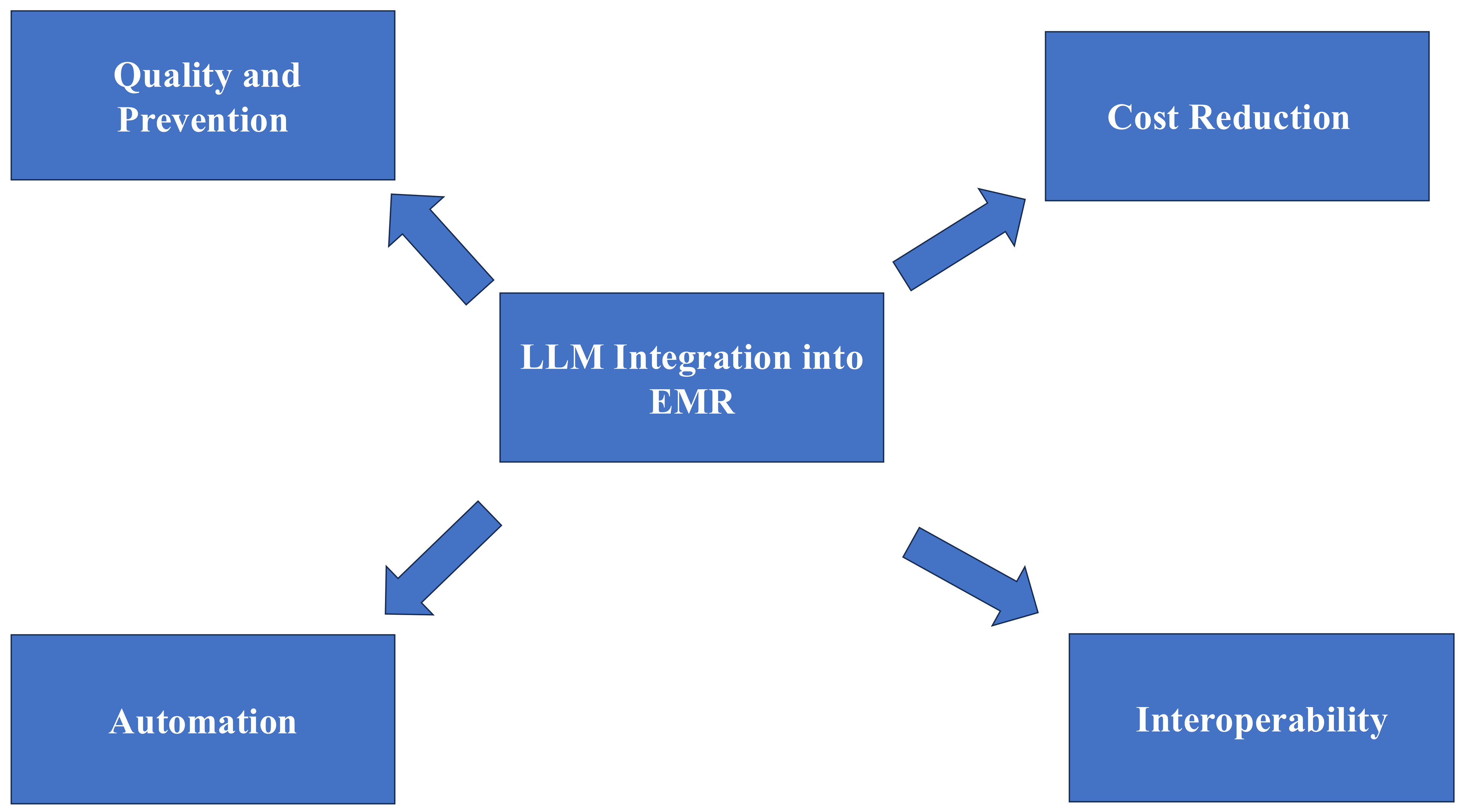
Key Insights:
- Leveraged GPT-4 Turbo to structure unstructured historical colonoscopy data.
- Achieved accurate, personalized recall intervals through AI-driven workflows.
- Identified and mitigated outliers in data to enhance reliability.

Impact:
- Prevented over **6092 potential colorectal cancer cases**.
- Significant economic benefits with an estimated **$400M (initial care)–$671M (end of life) in treatment cost savings**.
- Enhanced clinician confidence with transparent, comment-tagged recall recommendations.

Lessons Learned:
- Continuous prompt refinement and stakeholder collaboration are vital.
- Scalability requires robust quality control and filtering mechanisms.
- Regular updates to workflows ensure sustained accuracy and usability.

Future Directions:
- Collect ongoing clinician feedback for continuous improvement (e.g., address satellite clinic data integration for comprehensive deployment).
- Assess role of continuous integration of LLM into Epic to actively update patient colonoscopy recall data.
- Consider expansion of the model to other preventive care areas (e.g., mammography).

Figure 3. Impact of LLM Integration on Health Maintenance



Clusmann *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141. 2023.
National Cancer Institute. Financial burden of cancer care. Cancer Trends Progress Report. Reviewed March 2024. Accessed July 2024.
Jodal et al. Long-term colorectal cancer incidence and mortality after adenoma removal in women and men. *Aliment Pharmacol Ther*. 2022