Leveraging Machine Learning and Robotic Process Automation to identify and convert unstructured colonoscopy results into actionable data: a proof of concept

Elizabeth R Stevens, Jager Hartman, Paul Testa, Ajay Mansukhani, Casey Monina, Amelia Shunk, David Ranson, Yana Imberg, Ann Cote, Dinesha Prabhu, Adam Szerencsy

Submitted to: Journal of Medical Internet Research
on: March 05, 2025

Table of Contents

Leveraging Machine Learning and Robotic Process Automation to identify and convert unstructured colonoscopy results into actionable data: a proof of concept

Elizabeth R Stevens[1, 2] PhD, MPH; Jager Hartman[2]; Paul Testa[2] MD; Ajay Mansukhani[2]; Casey Monina[2] RN; Amelia Shunk[1] MMCi; David Ranson[2]; Yana Imberg[2]; Ann Cote[2]; Dinesha Prabhu[2]; Adam Szerencsy[2] DO

[1]Department of Population Health Grossman School of Medicine New York University New York US
[2] Medical Center Information Technology NYU Langone Health New York US

Corresponding Author:
Elizabeth R Stevens PhD, MPH
Department of Population Health
Grossman School of Medicine
New York University
227 E30th St
Rm 636
New York
US

Abstract

Background: Effective Colorectal cancer (CRC) screening is a cornerstone of preventive healthcare. Existing electronic health record (EHR) tools to facilitate reminders for CRC screening follow-up are inadequate to address clinician needs. With rising patient volumes and a focus on quality, our health system had the objective to create a more efficient way to ensure accurate documentation of CRC screening follow-up intervals from inbound colonoscopy reports. We developed an integrated end-to-end workflow solution using machine learning (ML) and robotic process automation (RPA) to extract and update follow-up dates from unstructured data.

Objective: To automate data extraction from external, free-text colonoscopy reports to identify and document recommended follow-up dates for CRC screening in structured fields within the EHR.

Methods: As proof of concept, we outline the process development, validity, and implementation of an approach that integrates available tools to automate data retrieval and entry within the EHR of a large academic health system. The health system uses Epic Systems as its EHR platform. This proof-of-concept process study consisted of six stages: 1) identification of gaps in documenting recommendations for follow-up CRC screening from external colonoscopy reports; 2) defining process objectives; 3) identification of technologies; 4) creation of process architecture; 5) process validation; and 6) health system-wide implementation. Chart review was performed to validate process outcomes and estimate impact.

Results: We developed an automated process with three primary steps that leveraged ML and RPA to create a fully orchestrated workflow to update the CRC screening recall date based on colonoscopy reports received from external sources. Process validity was assessed with 690 scanned colonoscopy reports. From the organization-wide implementation go-live date until December 31st, 2024, the system has processed 16,563 external colonoscopy reports. Of these 35.3% (5841) had a follow-up date that met the relevant threshold by the ML model and thus were identified as ready for RPA processing. This resulted in an estimated increase in documentation accuracy of 27.2%.

Conclusions: The implementation of an automated workflow to extract and update CRC screening follow-up dates from external colonoscopy reports is feasible and has the potential to improve accuracy in patient recall based on recommendations while reducing clinician documentation burden. Expanding similar processes to other types of unstructured data could provide another mechanism to solve for a lack of data integration and improve reporting for quality measures within the EHR. Automated workflows leveraging ML and RPA offer practical solutions to overcome interoperability challenges and enhance the use of unstructured data within healthcare systems.

(JMIR Preprints 05/03/2025:73504)

DOI: https://doi.org/10.2196/preprints.73504

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Leveraging Machine Learning and Robotic Process Automation to identify and convert unstructured colonoscopy results into actionable data: a proof of concept

Elizabeth R. Stevens, PhD, MPH[1,2]; Jager Hartman; Paul Testa, MD[2,3]; Ajay Mansukhani[3]; Casey Monina, RN[3]; Amelia Shunk, MMCi[1]; David Ranson[3]; Yana Imberg [3]; Ann Cote[3]; Dinesha Prabhu[3]; Adam Szerencsy, DO[2,3,4]

1. Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA
2. Department of Health Informatics, New York University Langone Health, New York, NY, USA
3. Medical Center Information Technology, New York University Langone Health, New York, NY, USA
4. Department of Medicine, New York University Grossman School of Medicine, New York, New York, USA.

*Corresponding author:
Elizabeth R. Stevens, PhD, MPH
227 E30th St
Rm 636
New York, NY 10016, USA
+1-646-501-2558
elizabeth.stevens@nyulangone.org

Keywords: colorectal cancer; RPA; automation; machine learning; artificial intelligence

Word Count: 3919 / 4000
Tables: 0 / 4
Figures: 2 / 6
References: 34

ABSTRACT

Background: Effective Colorectal cancer (CRC) screening is a cornerstone of preventive healthcare. Existing electronic health record (EHR) tools to facilitate reminders for CRC screening follow-up are inadequate to address clinician needs. With rising patient volumes and a focus on quality, our health system had the objective to create a more efficient way to ensure accurate documentation of CRC screening follow-up intervals from inbound colonoscopy reports. We developed an integrated end-to-end workflow solution using machine learning (ML) and robotic process automation (RPA) to extract and update follow-up dates from unstructured data.

Objective: To automate data extraction from external, free-text colonoscopy reports to identify and document recommended follow-up dates for CRC screening in structured fields within the EHR.

Materials and Methods: As proof of concept, we outline the process development, validity, and implementation of an approach that integrates available tools to automate data retrieval and entry within the EHR of a large academic health system. The health system uses Epic Systems as its EHR platform. This proof-of-concept process study consisted of six stages: 1) identification of gaps in documenting recommendations for follow-up CRC screening from external colonoscopy reports; 2) defining process objectives; 3) identification of technologies; 4) creation of process architecture; 5) process validation; and 6) health system-wide implementation. Chart review was performed to validate process outcomes and estimate impact.

Results: We developed an automated process with three primary steps that leveraged ML and RPA to create a fully orchestrated workflow to update the CRC screening recall date based on colonoscopy reports received from external sources. Process validity was assessed with 690 scanned colonoscopy reports. From the organization-wide implementation go-live date until December 31st, 2024, the system has processed 16,563 external colonoscopy reports. Of these 35.3% (5841) had a follow-up date that met the relevant threshold by the ML model and thus were identified as ready for RPA processing. This resulted in an estimated increase in documentation accuracy of 27.2%.

Discussion: The implementation of an automated workflow to extract and update CRC screening follow-up dates from external colonoscopy reports is feasible and has the potential to improve accuracy in patient recall based on recommendations while reducing clinician documentation burden. Expanding similar processes to other types of unstructured data could provide another mechanism to solve for a lack of data integration and improve reporting for quality measures within the EHR. Automated workflows leveraging ML and RPA offer practical solutions to overcome interoperability challenges and enhance the use of unstructured data within healthcare systems.

Keywords: colorectal cancer; RPA; automation; colonoscopy; machine learning; artificial intelligence

BACKGROUND AND SIGNIFICANCE

Colorectal cancer (CRC) is the second most common cause of cancer related death in the United States [1] and has a growing incidence rate, particularly among younger individuals.[2] Effective CRC screening is a cornerstone of preventive healthcare, significantly reducing both the incidence and mortality rates associated with CRC.[3] To realize the full benefit of CRC screening, accurate and timely recall of patients for follow-up screening is needed.

Despite the availability of electronic health record (EHR) tools to facilitate reminders for CRC screening follow-up, these tools often fail to address clinician needs.[4, 5] The limited interoperability of health information exchanges (HIE) between healthcare systems, and reliance on unstructured data to convey recommendations for follow up, place a high documentation burden on clinicians, leading to substantial missed opportunities for improved CRC follow-up screening and surveillance.[6-8]

EHR tools are rarely responsive to patient-specific factors or colonoscopy and pathology findings. Consequently, recalling a patient for more frequent follow-up screening requires a manual update.[9] To enhance CRC screening follow-up accuracy, while minimizing documentation burden, solutions that automate the retrieval and processing of colonoscopy results are needed. An automated process for updating CRC screening follow-up based colonoscopy findings is likely to increase the number of patients with a documented guideline-concordant CRC follow-up screening interval.[10][11]

As with nearly 80% of EHR data,[12] colonoscopy and pathology findings are frequently documented within the EHR as free text.[13] Further complexity arises when colonoscopy results are received as inbound medical records from external sources. These records, often received via fax, mail, or email, require additional steps to document the relevant data into the her, posing a significant administrative burden for documenting CRC screening due dates.[14]

Natural language processing (NLP) models have the potential to address barriers associated with leveraging unstructured data and accurately extract relevant data from free-text colonoscopy and pathology reports.[15] However, NLP alone cannot address all workflow steps required to convert text-based reports into structured fields in the EHR. It is essential to pair the model with supplemental tools for data entry. A solution is needed that orchestrates multiple technologies to address each step in the workflow necessary to update the appropriate CRC follow-up dates.

With rising patient volumes and a focus on quality, our health system needed a more efficient way to ensure accurate documentation of CRC screening follow-up intervals from inbound colonoscopy reports. To address this, we developed an integrated workflow using machine learning (ML) and robotic process automation (RPA) to extract and update follow-up dates from unstructured data.

OBJECTIVE

As proof of concept, we outline the process and validity of this approach, which integrates available tools to automate data retrieval and entry within the EHR. This framework can be adapted to other clinical information to improve reporting, data accuracy, and workflow efficiency.

MATERIALS AND METHODS

This proof-of-concept study was part of an institutional initiative to improve the accuracy of documented CRC screening follow-up recommendations extracted from unstructured colonoscopy reports from external organizations. The study developed an integrated workflow using existing technologies to address this care gap and assessed its validity. This study consisted of six stages: 1) identification of gaps in operationalizing recommendations for follow-up CRC screening from external colonoscopy reports; 2) defining process objectives; 3) identification of available technological tools; 4) creation of process architecture; 5) process validation and refinement; and 6) health system-wide process implementation.

As an operational quality improvement project, the New York University Langone Health (NYULH) Institutional Review Board (IRB) waived approval for the study procedures. A waiver of HIPAA authorization was included to access patient medical records. All study procedures complied with institutional ethical standards.

Study Setting and Participants

The NYULH system is a private, non-profit hospital system serving the greater New York area and Florida with over 9 million patients. The NYULH system uses Epic Systems as its EHR platform. For the process validation component of this study, EHR data were extracted for a series of four sequential simple random samples of colonoscopy reports from adult patients aged 45-75 years from external sources between August and September 2023.

Health Informatics Care Gap

Ensuring timely follow-up for CRC screening was identified as an important quality measure for the health system. Accurately tracking patients due for CRC screening is facilitated by having the next colonoscopy follow-up date documented discretely in the EHR’s Health Maintenance (HM) activity.

Gaps in this process frequently occur when colonoscopy reports from external organizations are scanned into the health system's EHR. Since these documents contain unstructured data, critical information remains inaccessible unless manually transcribed into the EHR.[12, 13] If this information is not updated manually, the HM follow-up screening interval defaults to ten years, which may not be appropriate for many patients.[8] As a result, the HM activity often fails to reflect the recommended follow-up intervals indicated in the colonoscopy report unless users manually update it. Factors such as procedural findings, pathology results, and patient-specific risk factors may necessitate more frequent surveillance. To ensure transparency and awareness for both patients and clinicians, as well as for accurate reporting, this information should be documented discretely in the appropriate fields within the EHR.

Manually updating the HM follow-up intervals is time-consuming, often leading to unchanged follow-up dates in the EHR and failing to reflect the appropriate recall time frame. This documentation gap can result in missed or delayed surveillance, potentially impacting patient outcomes. To address this challenge, an efficient and effective process is needed to extract and process unstructured text from colonoscopy reports to automatically update HM follow-up intervals.

Defining the Process Objective

After identifying the care gap, the objective was to develop a process to automate updating HM based on the external colonoscopy reports. We sought to leverage ML to identify follow-up recommendations within the reports and RPA to automate the entry of recommended recall dates into the HM activity.

It was also determined that the process must operate within the following constraints: 1) Use existing technologies within the health system; 2) Not override any recommendations that have been manually entered by a clinician; 3) ML engine identifies follow-up date above a technical performance threshold of >70% confidence; and 4) No human intervention after the colonoscopy report has been scanned into the EHR.

Although pathology results should ideally serve as the gold standard for assessing follow-up intervals, due to limited availability, the recommendation listed within the colonoscopy report was deemed a sufficiently accurate alternative.

Orchestrated Workflow Development

To develop the orchestrated workflow, the team first broke the process into concrete steps. We then mapped out existing technological components within the health IT ecosystem needed for each action. The steps included: 1) Daily extraction of scanned external colonoscopy reports; 2) A mechanism for an ML model to receive and analyze those reports to identify the next follow-up date; and 3) An automated process that updates the HM activity when appropriate.

The team established care assumptions to guide when the process updates would take precedence over existing HM follow-up frequencies. These assumptions included: 1) All updates must be associated with a colonoscopy order; 2) The bot should only update HM if the scanned report is more recent than the last documented colonoscopy; 3) The scanned report would not override manual entry; 4) The bot should not update a follow-up date if the patient has an existing earlier follow-up date; 5) If the ML engine's determination of follow-up date is below the performance threshold, the report should be queued for manual review; 5) If an outstanding colonoscopy order exists and the procedure date is after the order date, the report should be scanned to that order; 6) If there is a range of follow up dates (e.g., return in 3-5 years) the bot will use the earlier date. Based on these assumptions, business exceptions within the decision rules were defined to establish scenarios in which it would be inappropriate for the bot to update the HM.

Identified Technological Tools

Existing technologies were evaluated for their applicability to each workflow step, resulting in a process comprising six key components, including: 1) An internally developed ML model; 2) Fax management software; 3) A document management platform; 4) Input spreadsheets; 5) RPA; and 6) Epic's HM activity.

Fax Management Software

RightFax is an enterprise fax server software that allows the organization to send and receive faxes. It also transforms faxes for easier management and integration into other digital systems, including the EHR.[16] RightFax was used to receive colonoscopy reports from external sources.

Machine Learning (ML)

ML techniques, including Optical Character Recognition (OCR) and NLP, enable systems to "read" documents and extract relevant information.[17, 18] These algorithms can identify patterns, structures, and key data within various document formats, automating the extraction process and reducing the need for manual intervention. This capability is widely used in data entry, document analysis, and information retrieval. A proprietary model was internally developed to identify the follow-up dates for the next screening. This process required model training by a clinical subject matter expert.

Robotic Process Automation (RPA)

UiPath RPA is a technology that uses software robots, or "bots," to automate repetitive and rule-based tasks typically performed by humans. These bots can mimic human actions such as data entry, processing transactions, and managing records, thereby increasing efficiency and reducing errors. Unlike artificial intelligence, RPA follows predefined workflows and does not learn or adapt on its own.[19] In this case, the RPA tool "read" the output of the ML engine, updated the HM activity date if applicable, and documented the actions taken. If any business exceptions existed, they were similarly documented. Finally, the RPA sent a summary of the completed process to operational leadership for review.

Document Management Platform

OnBase is an enterprise content management platform developed by Hyland Software. It helps organizations manage documents in a unified system, acting as a digital filing cabinet with advanced automation capabilities. OnBase supports document storage, quick retrieval, and task automation, enhancing efficiency and compliance.[20] In this process, OnBase is used as the underlying application where the reports were scanned into the EHR. A specific directory was created to process incoming colonoscopy reports so that they can be processed by the ML model.

Input Spreadsheet

The input spreadsheet was an excel document that was used as a working tool to list the findings and follow up dates of the colonoscopy reports as identified by the ML model. Any findings with low confidence were indicated as such. The RPA solution reviewed the spreadsheet, "read" the follow up dates, and document what actions it took.

Health Maintenance (HM)

Epic's HM activity is an EHR feature designed to help providers manage and track preventive care and chronic disease management for patients. It allows healthcare organizations to set up and monitor recommended screenings, vaccinations, and follow-up care based on clinical guidelines, patient demographics, and medical history. It serves as a designated location in the EHR to document completed colonoscopies and the recommended follow-up date. However, it is not linked to clinical notes and must be manually updated to reflect changes due to pathology findings or patient-specific conditions.[6]

Validation

Prior to systemwide activation, outcome validation was performed via manual review of a sample of colonoscopy reports and associated RPA actions in the EHR. Each colonoscopy report reviewed by trained clinical analysts (DR & CM) for accuracy. Discrepancies were reviewed with the physician informaticist lead (AS). The validation outcomes were divided into two phases: 1) ML evaluation, determining whether a follow-up date was identified within the colonoscopy report and ready for bot processing; and 2) RPA bot action documented in the EHR. Reports not ready for the bot were categorized as a) correct, with no follow-up date specified; b) incorrect, having a follow-up date listed that was not picked up by the ML; or c) inaccurate follow-up date identified. Outcomes for reports processed by the RPA update were categorized as: a) a correct HM update; b) correct business exception assignment; or c) incorrect HM update. Four rounds of validation testing were performed with a prospective sample of y reports manually reviewed until a saturation of automated outcomes was observed.

Process Refinement

A protocol was established to modify and refine the ML algorithm and bot activities if error patterns were identified during validation. After each round of validation testing, if an error was addressable, the team adjusted either the ML model or RPA based on error source.

Governance Approval

Validation study findings were presented at leadership committee meetings to obtain governance approval for system-wide implementation. Besides confirming the process's validity, the approval process emphasized potential improvements in Health Maintenance follow-up date accuracy compared to the current reliance on a default 10-year interval and limited manual updates by clinicians.

Process implementation

After validation, refinement, and all necessary approvals, the process was implemented in the live environment. This began with a period of careful monitoring of the process known as hypercare. [21] The automated process was first implemented in the live environment as small batches, where ten reports were run through the process and assessed for validity before proceeding to the next ten. This procedure was performed until it was determined that the process performed appropriately in the live environment. After the initial phase, the process was expanded to analyze all prospectively received external colonoscopy reports.

RESULTS

Our system established a streamlined, three-step process integrating preexisting technologies to automate updates to the CRC screening follow-up date based on external colonoscopy results. The process steps, the workflow, and technology resources used for each function are presented in Figure 1.

1. Document Identification and Routing: Scanned colonoscopy reports are identified and sent to a designated work queue in the OnBase shared drive.
2. Machine Learning (ML) Processing: An ML model verifies that the document is a colonoscopy report, then extracts key data, including patient ID, report date, and follow-up recommendations. If data meets predefined accuracy thresholds, it is added to an input spreadsheet and labeled "RPA ready." If below the threshold, the document is marked “RPA not ready” and sent to a work queue for human validation.
3. Robotic Process Automation (RPA) Execution: The bot processes "RPA ready" entries by accessing the patient’s EHR, comparing the extracted follow-up date with existing HM follow-up frequency, and updating recall date if no business exceptions are detected. All processed data, including business exceptions, are documented in the input spreadsheet.

Business Exceptions

Nine business exceptions were defined such that if any criteria were met, the bot would not update HM. These criteria included: 1) The follow up date in the scanned document is longer than existing HM follow-up frequency; 2) A patient has a HM frequency modifier that sets a patient-specific screening frequency; 3) The next HM due date has been manually assigned after the scanned procedure date. These exceptions ensure the bot never takes precedence over a provider-set screening frequency or patient-specific factors; 4) The report has a follow-up date over 10 years; 5) The input spreadsheet cannot assess follow up date; 6) The procedure date linked to that document does not match the scanned document date; 7) The patient does not have the colonoscopy screening topic assigned in Epic due to age criteria; 8) The patient is not found; or 9) An order for the scanned colonoscopy report is not found. This exception ensures that the HM updates are associated with an appropriate colonoscopy order.

Validation Outcomes

Figure 2 presents the validation outcome findings. Process validity was assessed for 690 scanned colonoscopy reports. Of these, 425 (61.6%) were categorized by the ML algorithm as not ready for the bot. Of those, 130 (30.6%) listed a follow-up date not detected by the ML algorithm and 295 (69.4%) did not specify a follow-up date. Among the 265 reports identified as ready for the bot, 188 (70.9%) HM were correctly updated, 74 (27.9%) were flagged with an update exception, and 3 (1.1%) HM were incorrectly updated. The update automation process increased HM accuracy by 27.2%.

Process Refinements

During validation, the three incorrectly updated HM were due to (1) the bot creating an update based on an older report and (2) overriding a provider modification. These errors led to process refinement by updating the RPA steps to validate clinical information. The ML error rate of 18.8% for missing an available follow-up date was deemed acceptable, as refining the model to be more sensitive led to extraction of incorrect dates, seen as a greater issue than missing a date.

System-wide Implementation

The solution was implemented system-wide on October 4th, 2023. From go-live until

December 31st, 2024, the system processed 16,563 external colonoscopy reports. Of these, 35.3% (5841) had a follow-up date meeting the ML model threshold and thus identified as RPA Ready. Of the RPA Ready documents, the HM topic was successfully updated in 4,512 patient records. Among all processed reports, 72.8% (12,051) had business exceptions and were not updated; this included 22.8% (1,329) that were RPA Ready. Of the RPA Ready documents with business exceptions, 3.9% (225) had a patient-specific follow-up date modifier, 3.2% (189) had a more recent procedure, 6.8% (398) were manually updated by a provider, 8.3% (485) were outside standard screening age and did not have a colonoscopy HM topic assigned, and 0.6% (32) had a discrepancy where the procedure result date did not match the scanned document.

DISCUSSION

This proof-of-concept demonstrated the feasibility of using ML and RPA to develop an automated process capable of extracting key clinical information from unstructured documents and converting it into discrete, accessible data. The system improved the accuracy of CRC screening follow-up dates while reducing the documentation burden on clinical staff. This pragmatic and innovative solution represents one of the first end-to-end solutions to address the challenges of processing faxed, text-based results by converting free text into structured data. This work provides a template for healthcare systems to reduce documentation burden, improve follow-up date accuracy, and address care gaps arising from a lack of system integration.

The process automates all workflow stages from the initial receipt of the free text report to the entry of the follow-up date into the EHR. Traditionally, this information would be difficult to locate within existing workflows, limiting its influence on patient recall processes. By structuring and integrating this data, the system ensures timely follow-ups and supports downstream workflows to recall patients within recommended timeframes, ultimately improving preventive care management.

Many important data elements are not updated in the her due to their time-consuming nature.[22] This strategy reduces documentation burden while minimizing data-entry errors. As demonstrated in the validation study, this process increased HM CRC screening date accuracy rate by nearly 30%, contributing to important health system quality measures. This accuracy increase has the potential for a substantial downstream impact on patient outcomes.

Beyond improving follow-up date accuracy, this workflow reduces the administrative burden on clinical staff, ensuring timely and appropriate care. Transforming unstructured text into structured data also has significant implications for quality reporting. By automating data extraction and integration into the EHR, healthcare systems can more effectively track completed screenings and tests, ensuring comprehensive reporting on quality measures, closing care gaps, and improving population health management.

The use of automation, including RPA, has been previously proposed as a theoretical solution to address limitations in the EHR and other elements of data-driven healthcare.[23-27] However, prior literature has not presented a fully integrated end-to-end automated workflow. Prior research has primarily focused on data automation in structured formats[27] such as healthcare claims[25, 26] and the extraction and processing of EHR data for research.[27-32] A few studies present cases for using ML and AI for other data types,[30, 32] but their full integration into health system operations remains hypothetical. Our work builds upon previous research to demonstrate the feasibility and validity of implementing a pragmatic, automated workflow within the EHR to address real-world challenges. While we developed a proprietary model to extract follow-up dates from colonoscopy reports, we anticipate advancements in generative AI will enable large language models to achieve similar outcomes. These models can support this use case and be scaled for extracting other critical clinical data from unstructured text, reducing documentation burden and enhancing automation in healthcare systems.

This process framework can address inefficiencies from the lack of EHR interoperability or discordant information within various areas of the EHR. Our institution has adapted the described process to automate the documentation of follow-up dates for internally performed colonoscopies documented within ancillary systems and interfaced back as free text reports. Before a recent (2023) institutional effort to code colonoscopy pathology findings,[11] all colonoscopy reports were stored as unstructured text. Using a modified version of this automated process the institution has retrospectively processed 80,864 internal reports from May 2020 to April 2023 updating 71.8% (58,060) CRC screening HM follow-up dates. This successful application of ML combined with RPA highlights their potential to address other high-burden, low-complexity tasks in the EHR and could address deficiencies in structured EHR documentation. Scaling similar processes to other report types with discrete values, such as cholesterol or A1c, may demonstrate an even greater success.

Incorporation ML into the workflow was critical to the automated process' utility. Nearly 80% of EHR data,[12] including colonoscopy reports and pathology findings, are documented as free text.[13] Therefore, the critical information within the reports remains inaccessible in the EHR unless manually transcribed.[12, 13] Most proposed automations cannot use free-text data and rely on discretely stored data, such as laboratory result values,[27] which is unavailable for many EHR data sources unless data entry systems and clinician behaviors are modified. By documenting discrete data without requiring behavior change or system structure modifications, these tools may make automating a broader range of data capture more feasible.

Limitations

This study had several limitations. First, as this workflow integrates multiple disparate technologies to create an automated process, it is susceptible to technical issues from changes to any of the individual components. For example, system updates can change the user interface, causing the bot to no longer recognize its inputs. To address this, the process requires consistent monitoring and maintenance. Second, using colonoscopy reports rather than pathology results as the source for follow-up interval means some dates may require updating after pathology results are available. However, as colonoscopy reports are more commonly transmitted from external sources, this decision was pragmatic and provided a sufficiently accurate representation to process the greatest number of results. Future iterations may consider incorporating logic based on the pathology results if included along with the colonoscopy report. Finally, as more types of report formats are received, regular refinement of the ML model is needed to increase the ability of the model to identify follow-up dates. However, due to finite resources, the accuracy deemed sufficient for operational purposes, and available backup manual processes, our model is not regularly refined. Moreover, regardless of ML refinement, the process's effectiveness was limited by the high rate of missing elements in the received colonoscopy reports.[33, 34] This should be considered when building automated processes intended to entirely replace manual supervision.

CONCLUSION

This study demonstrates the feasibility of developing a high-validity, automated process to update CRC screening follow-up recommendations within the EHR using scanned colonoscopy reports. By leveraging ML and RPA, we successfully extracted key clinical information from unstructured documents and converted it into structured, actionable data. The success of this approach highlights the potential for similar automation strategies in other high-volume, low-complexity documentation tasks in healthcare, enhancing efficiency, reducing manual errors, and improving patient outcomes. As AI advances, integrating more sophisticated large language models may further refine and scale these efforts. Future work should evaluate the long-term impact, scalability across different healthcare systems, and additional opportunities for automation to support clinical workflows, quality reporting, and patient care.

ACKNOWLEDGEMENTS

Author contributions: All authors meet the criteria for authorship stated in the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*. The specific areas of contributions are as follows: study concept and design (ERS,PT, AM, A Szerencsy); Acquisition of data (CM, DR, YI, AC, DP); analysis and interpretation of data (all authors); Drafting of the manuscript (ERS, A Shunk, A Szerencsy); and critical revision of the manuscript for important intellectual content (all authors).

Competing Interests: The authors have no competing interests to declare.

Funding: This work was funded by operational funds from NYU Langone Health.

Data availability: The data underlying this article cannot be shared publicly due to privacy of clinical data used in the study. The data will be shared on reasonable request to the corresponding author.

AI disclosure: AI was not used in the writing of this manuscript.

REFERENCES

1. Dharwadkar P, Zaki TA, Murphy CC. Colorectal Cancer in Younger Adults. Hematol Oncol Clin North Am. 2022 Jun;36(3):449-70. PMID: 35577711. doi: 10.1016/j.hoc.2022.02.005.
2. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA: A Cancer Journal for Clinicians. 2023 2023/05/01;73(3):233-54. doi: https://doi.org/10.3322/caac.21772.
3. Ladabaum U, Dominitz JA, Kahi C, Schoen RE. Strategies for Colorectal Cancer Screening. Gastroenterology. 2020 Jan;158(2):418-32. PMID: 31394083. doi: 10.1053/j.gastro.2019.06.043.
4. Atlas SJ, Tosteson ANA, Burdick TE, Wright A, Breslau ES, Dang TH, et al. Primary Care Practitioner Perceptions on the Follow-up of Abnormal Cancer Screening Test Results. JAMA Network Open. 2022;5(9):e2234194-e. doi: 10.1001/jamanetworkopen.2022.34194.
5. Makam AN, Lanham HJ, Batchelor K, Samal L, Moran B, Howell-Stampley T, et al. Use and satisfaction with key functions of a common commercial electronic health record: a survey of primary care providers. BMC Medical Informatics and Decision Making. 2013 2013/08/09;13(1):86. doi: 10.1186/1472-6947-13-86.
6. Gaffney A, Woolhandler S, Cai C, Bor D, Himmelstein J, McCormick D, et al. Medical Documentation Burden Among US Office-Based Physicians in 2019: A National Study. JAMA internal medicine. 2022;182(5):564-6. doi: 10.1001/jamainternmed.2022.0372.
7. Johnson KB, Stead WW. Making Electronic Health Records Both SAFER and SMARTER. JAMA. 2022;328(6):523-4. doi: 10.1001/jama.2022.12243.
8. Øines M, Helsingen LM, Bretthauer M, Emilsson L. Epidemiology and risk

factors of colorectal polyps. Best Practice & Research Clinical Gastroenterology. 2017 2017/08/01/;31(4):419-24. doi: https://doi.org/10.1016/j.bpg.2017.06.004.
9. Coronado GD, Burdick T, Petrik A, Kapka T, Retecki S, Green B. Using an Automated Data-driven, EHR-Embedded Program for Mailing FIT kits: Lessons from the STOP CRC Pilot Study. J Gen Pract (Los Angel). 2014 Jan 5;2. PMID: 25411657. doi: 10.4172/2329-9126.1000141.
10. Schrembs TR. Use of electronic health record reminders to improve primary care providers' colorectal cancer screening recommendations. Doctor of Nursing Practice Papers. 2023;Paper 130.
11. Stevens ER, Nagler A, Monina C, Kwon J, Olesen Wickline A, Kalkut G, et al. Pathology-Driven Automation to Improve Updating Documented Follow-Up Recommendations in the Electronic Health Record After Colonoscopy. Clin Transl Gastroenterol. 2024 Nov 13. PMID: 39665587. doi: 10.14309/ctg.0000000000000785.
12. Murdoch TB, Detsky AS. The inevitable application of big data to health care. Jama. 2013 Apr 3;309(13):1351-2. PMID: 23549579. doi: 10.1001/jama.2013.393.
13. Millares Martin P. Consultation analysis: use of free text versus coded text. Health and Technology. 2021 2021/03/01;11(2):349-57. doi: 10.1007/s12553-020-00517-3.
14. Ford B. Colonoscopy Follow-up: U.S. Multi-Society Task Force on Colorectal Cancer Updates Recommendations. Am Fam Physician. 2021 Mar 1;103(5):314-6. PMID: 33630559.
15. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol. 2013 Jun;11(6):689-94. PMID: 23313839. doi: 10.1016/j.cgh.2012.11.035.
16. Open Text Corporation. OpenText Fax (RightFax). 2024 [December 19, 2024]; Available from: https://www.opentext.com/products/fax.
17. Learn Microsoft. OCR - Optical Character Recognition. 2024; Available from: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr.
18. Fanni SC, Febi M, Aghakhanyan G, Neri E. Natural Language Processing. In: Klontzas ME, Fanni SC, Neri E, editors. Introduction to Artificial Intelligence. Cham: Springer International Publishing; 2023. p. 87-99.
19. IBM. What is robotic process automation (RPA)? 2024 [October 1, 2024]; Available from: https://www.ibm.com/topics/rpa.
20. Hyland. What can you do with OnBase? 2024 [October 1, 2024]; Available from: https://www.hyland.com/en/resources/articles/what-can-you-do-with-onbase.
21. Wren H. Hypercare: What it means and why it matters in CX. 2024 [December 19, 2024]; Available from: https://www.zendesk.com/blog/hypercare/.
22. Stevens ER, Nagler A, Monina C, Kwon J, Olesen Wickline A, Kalkut G, et al. Pathology-Driven Automation to Improve Updating Documented Follow-Up Recommendations in the Electronic Health Record After Colonoscopy. Clinical and Translational Gastroenterology. 2024;15(12).
23. Dagli MM, Ghenbot Y, Ahmad HS, Chauhan D, Turlip R, Wang P, et al. Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review.

Sci Rep. 2024 2024/11/05;14(1):26783. doi: 10.1038/s41598-024-77535-y.
24. Venigandla K. Integrating RPA with AI and ML for Enhanced Diagnostic Accuracy in Healthcare. Power System Technology. 2022;46(4).
25. Deepak V, Prabhu K, Srinivasan R. Advanced Data Integration Techniques for Healthcare Claims Processing: Leveraging AI and Automation to Streamline Information Flow. Journal of Artificial Intelligence Research and Applications. 2024 01/07;4(1):774-818.
26. Prabhu K, Ravi Kumar B, Bhavani K. Automating Data Integration for Healthcare Claims Processing: Enhancing Accuracy and Speed with Cloud-Based Solutions. Journal of Machine Learning for Healthcare Decision Support. 2024 02/04;4(1):247-90.
27. Tiase VL, Hull W, McFarland MM, Sward KA, Del Fiol G, Staes C, et al. Patient-generated health data and electronic health record integration: a scoping review. JAMIA Open. 2020;3(4):619-27. doi: 10.1093/jamiaopen/ooaa052.
28. Cheng AC, Banasiewicz MK, Johnson JD, Sulieman L, Kennedy N, Delacqua F, et al. Evaluating automated electronic case report form data entry from electronic health records. J Clin Transl Sci. 2023;7(1):e29. doi: 10.1017/cts.2022.514.
29. Parciak M, Suhr M, Schmidt C, Bönisch C, Löhnhardt B, Kesztyüs D, et al. FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. BMC Medical Informatics and Decision Making. 2023 2023/05/15;23(1):94. doi: 10.1186/s12911-023-02195-3.
30. Muneer Ahmed S. Data Integration: AI-driven approaches to streamline data integration from various sources. Journal of AI-Assisted Scientific Discovery. 2023 03/08;3(1):668-94.
31. Kaspar M, Fette G, Hanke M, Ertl M, Puppe F, Störk S. Automated provision of clinical routine data for a complex clinical follow-up study: A data warehouse solution. Health Informatics Journal. 2022;28(1):14604582211058081. PMID: 34986681. doi: 10.1177/14604582211058081.
32. Pulivarthy P. Enhancing Data Integration in Oracle Databases: Leveraging Machine Learning for Automated Data Cleansing, Transformation, and Enrichment. International Journal of Holistic Management Perspectives. 2023 06/07;4(4):1-18.
33. Sharma RS, Rossos PG. A Review on the Quality of Colonoscopy Reporting. Can J Gastroenterol Hepatol. 2016;2016:9423142. PMID: 27446877. doi: 10.1155/2016/9423142.
34. Keswani RN, Crockett SD, Calderwood AH. AGA Clinical Practice Update on Strategies to Improve Quality of Screening and Surveillance Colonoscopy: Expert Review. Gastroenterology. 2021;161(2):701-11. doi: 10.1053/j.gastro.2021.05.041.

Figure 1. Automation process and technology tools used in each process step

Note: RPA-robotic process automation; HM-health maintenance

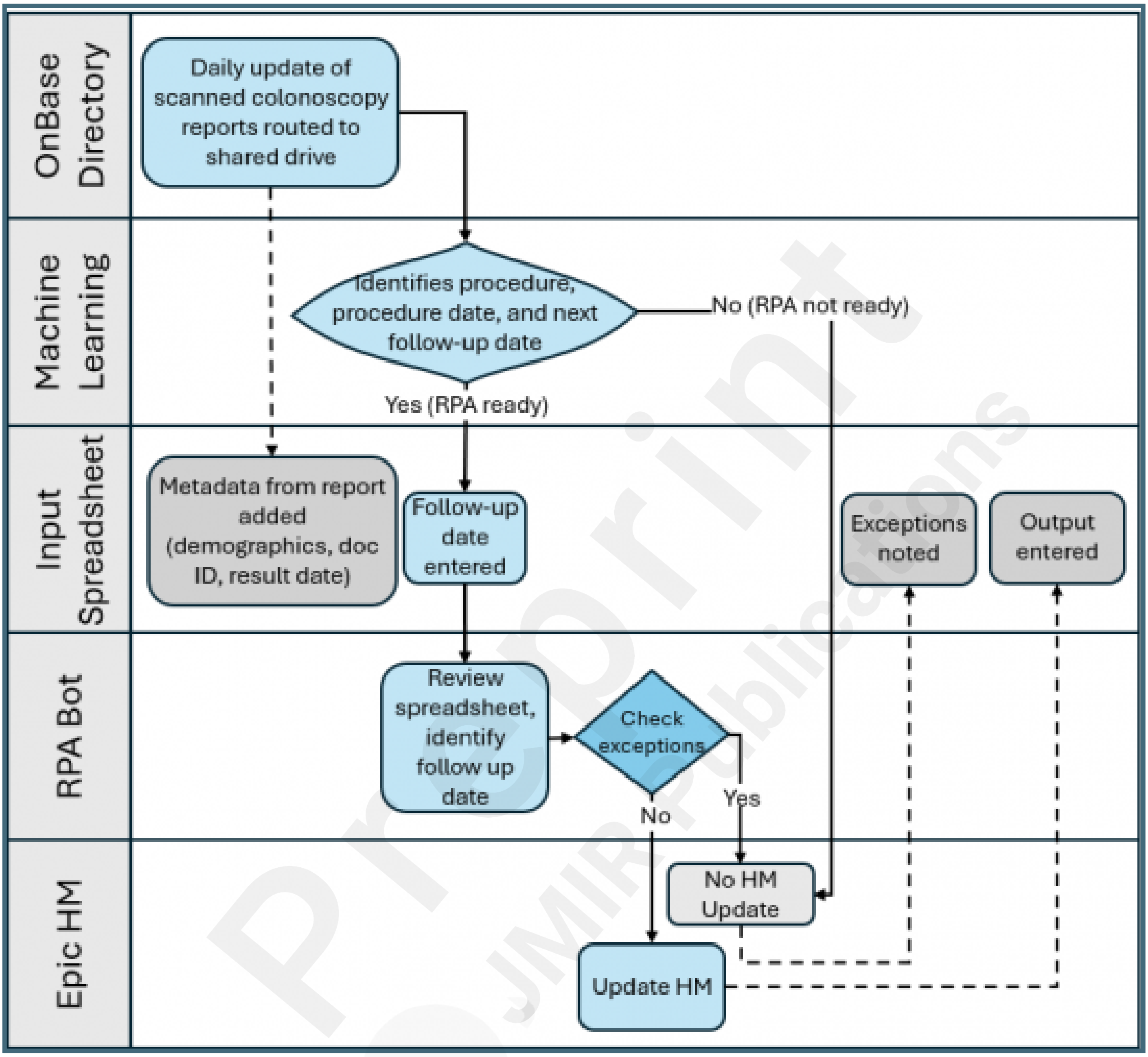
Figure 2. Process validation outcomes

Note: ML-machine learning; RPA-robotic process automation; HM-health maintenance

Supplementary Files

Figures

Automation process and technology tools used in each process step.



Process validation outcomes.