ORIGINAL RESEARCH

Electronic Health Record Algorithm Development for Research Subject Recruitment Using Colonoscopy Appointment Scheduling

Jeanette M. Daly, RN, PhD, Kim Parang, MA, and Barcey T. Levy, PhD, MD

Introduction: Electronic health records (EHRs) are often leveraged in medical research to recruit study participants efficiently. The purpose of this study was to validate and refine the logic of an EHR algorithm for identifying potentially eligible participants for a comparative effectiveness study of fecal immunochemical tests (FITs), using colonoscopy as the standard.

Methods: An Epic report was built to identify patients who met the eligibility criteria to recruit patients having a screening or surveillance colonoscopy. With the goal of maximizing the number of potentially eligible patients that could be recruited, researchers, with the assistance of information technology and scheduling staff, developed the algorithm for identifying potential subjects in the EHR. Two validation methods, descriptive statistics and manual verification, were used.

Results: The algorithm was refined over 3 iterations leading to the following criteria being used for generating the report: Age, Appointment Made On/Cancel Date, Appointment Procedure, Contact Type, Date Range, Encounter Departments, ICD-10 codes, and Patient Type. Appointment Serial Number/ Contact Serial Number were output fields that allowed the tracking of cancellations and reschedules.

Conclusion: Development of an EHR algorithm saved time in that most individuals ineligible for the study were excluded before patient medical record review. Running daily reports that included cancellations and rescheduled appointments allowed for maximum recruitment in a time frame appropriate for the use of the FITs. This algorithm demonstrates that refining the algorithm iteratively and adding cancellations and reschedules of colonoscopies increased the accuracy of reaching all potential patients for recruitment. (J Am Board Fam Med 2021;34:49–60.)

Keywords: Algorithms, Appointments and Schedules, Colonoscopy, Colorectal Cancer, Early Detection of Cancer, Electronic Health Records, Information Technology, International Classification of Diseases, Occult Blood

Introduction

Subject recruitment methods vary according to the research topic and population being studied, with advertisements, invitation letters, review of patient appointment lists, and/or electronic health records (EHRs) as commonly used methods. EHRs are often leveraged in medical research to recruit study participants efficiently affording cost containment and study success.[1,2] Medical diagnoses using ICD-10 codes are commonly used for subject recruitment via EHRs. More involved methods of identifying patients include searching medication lists, prescription data, or unstructured (ie, free-text) data when structured data elements do not exist.[3,4]

Well-programmed EHR algorithms have been found to enhance subject recruitment. In one study, researchers found recruitment of subjects was faster and more cost-efficient using the EHR patient portal when they used an algorithm to search for diagnosis codes and the medication list.[2] In another study, researchers compared the efficiency and enrollment rate of manual chart review versus an automated prescreening method using an algorithm for the recruitment of patients with the presence of

the following: a diagnosis of diabetes, high glucose levels, and a recent insulin order.[5] Their algorithm pulled from their data warehouse and not from their live electronic medical record, but it still significantly increased the number of subjects screened and enrolled.[5] Yet other researchers studying use of appropriate medications for individuals with asthma recruited patients through scheduled clinic visits noted in the EHR.[4] Throughout the recruitment process, they encountered problems with their algorithm, such as difficulty detecting same-day visits and clinic appointments that were cancelled or no-shows. Unable to remedy the problem, staff was assigned to review the EHR every day to look for appropriate appointments.[4] Investigators using the EHR for subject recruitment tend to use disease-oriented data; it is less common to use administrative data such as scheduling events.[4]

Our study was funded by the National Cancer Institute to address the knowledge gap in test characteristics of fecal immunochemical tests (FITs) using colonoscopy as the gold standard (referred to as the BestFIT study). Subjects were required to complete all the steps of the BestFIT study before their colonoscopy. A major problem with patients scheduled for a colonoscopy is that they cancel the procedure, may or may not reschedule, and are no-shows. This is different from a cancellation or no-show in a primary care clinic where patients are overbooked, since procedure suites can only schedule a finite number of patients. Patient no-show rates vary by study—for predominately African American populations, no-show rates have been 20%[6] and 23%;[7] in 23 small and large urban primary care physician offices a no-show rate of 38% was found for first-time colonoscopies;[8] in a large safety net health care system nonattendance was 42%.[9] To complete recruitment for the study in a timely manner and reduce the cost spent on recruiting patients to only have them not complete their colonoscopy, researchers needed to find a way to track patients' colonoscopy appointment statuses. An EHR algorithm could do this and generate a list of patients scheduled for a colonoscopy meeting specific inclusion and exclusion criteria. The purpose of this study was 2-fold: 1) to validate and refine the logic of a newly created EHR algorithm for identifying potentially eligible patients for the BestFIT study recruitment, and 2) to illuminate the possible pitfalls in constructing an EHR algorithm.

Methods

Institutional Review Board approval was obtained from the University of Iowa. This study was conducted in the Department of Family Medicine in an academic center with one clinic on campus and 4 clinics off-campus serving 112,000 individual patients of all ages. Through the Digestive Health Center, approximately 3600 screening and surveillance colonoscopies are performed each year. With the goal of maximizing the number of patients that could be recruited, researchers and information technology (IT) staff developed the rules for identifying potential subjects in the Epic EHR software, using the Epic Reporting Workbench, which allows users to pull data in real-time (Epic Systems, Verona, WI). The process with IT staff started during a pilot study 7 months before when the BestFIT study was funded.

To be eligible for the BestFIT study, patients had to be 50 to 85 years of age scheduled for a screening or surveillance colonoscopy. Those scheduled for a screening colonoscopy were asymptomatic patients testing for the presence of colorectal cancer or polyps having no history of colon cancer, polyps, and/or gastrointestinal disease. Those scheduled for a surveillance colonoscopy were asymptomatic patients at an interval less than the standard 10 years from the last colonoscopy, due to personal findings of cancer, polyps, or gastrointestinal disease on a previous examination. Seventy-one percent of the subjects were scheduled for screening colonoscopy and 29% for surveillance colonoscopy. Patients with familial polyposis syndromes, ulcerative colitis, Crohn's disease, personal history of colorectal cancer, or active rectal bleeding were to be excluded; patients with para- or quadriplegia, dementia, or severe psychiatric issues were also excluded. These requirements were expressed in the Epic algorithm through "criteria," which are the eligibility filters of the data query. In addition to the Age and ICD-10 code criteria described above, administrative criteria denoting appointment scheduling events, type of procedure, place of procedure, and date appointment was made were used to identify appropriate patients (Table 1).

The second component of the Epic algorithm was a display of the search results, whose variables are exported as a Microsoft Excel comma-separated values file (referred to as Output). The Output included patient medical record number, demographic information, and information pertaining to

Table 1. Three Iterations of FIT Daily Report Criteria—Definitions (D), Parameters (P), and Rationales (R)

Criteria			
First Iteration	Second Iteration	Third Iteration*	Definitions, Parameters, and Rationales
Age	Age	Age	D: Age at the time of making the appointment.
			P: Range for patients 'greater than or equal to 50 years and less than or equal to 85 years.'
Appointment Procedure			D: Procedure code designating the specific procedure to be included.
			P: Set to equal the scheduling code for the general global colonoscopy procedure ENDO COLONOSCOPY [1685], i.e., 'equal to 1685.'
			R: Used the one procedure code IT provided.
	Appointment Procedure (revised)	Appointment Procedure	P: Added more colonoscopy procedure codes to include 11 discrete codes for the study.
			R: Initially unaware there were 21 procedure codes, used additional appropriate codes to capture all eligible patients.
Appointment Status			D: The patient's appointment status, e.g., scheduled, cancelled, arrived, was bumped.
			P: Set as 'equal to Scheduled.'
			R: Initially unaware could capture appointment changes, such as cancelled.
	Appointment Status (deleted)	n/a	R: FIT Daily Report would output duplicates of the same appointments every time the report was run if the 'Scheduled' appointment fell within the specified Date Range. Replaced with Appointment Made on Date and Contact Type (below).
Date Range			D: Range of days from which the algorithm searches for colonoscopy appointments.
			P: Set as 'from W + 3 to W + 12' (21 to 84 or 85 days out from when the FIT Daily Report was run).
			R: This date range was set to send an invitation to eligible patients so stool specimen could be obtained before colonoscopy prep and within 4 months of scheduled colonoscopy.
	Date Range (revised)		P: Set as 'from W + 4 to W + 8' (29 to 56 days out from when the FIT Daily Report was run).
			R: The recruitment start timeframe was extended to allow for subject and study mailings.
		Date Range (revised)	D: Range of days from which the algorithm searches for colonoscopy appointment. Set as 'from T-1 to Y + 4' (from the day before the FIT Daily Report was run to 4 years out, the maximum allowed upper bound in Epic Reporting Workbench).
			P: The second iteration of the criterion missed appointments that were rescheduled beyond 56 days. Expanding the interval to encompass all reasonable dates on which the colonoscopies could be scheduled ensured that all scheduling changes could be tracked by researchers.
			R: Enforcing the recruitment timeframe (where colonoscopy appointment is between W + 4 and W + 8) was relegated to the Tracking Database.
Encounter Department			D: Code identifying the department and location of the room where the patient would encounter the colonoscopy.
			P: Set as 'equal to Med GI/Hepatology Procedure Unit [10401129].'
			R: IT provided one procedure unit code.

Continued

Table 1. Continued

	Criteria		
First Iteration	Second Iteration	Third Iteration*	Definitions, Parameters, and Rationales
	Encounter Department (revised)	Encounter Department	P: Expanded to include a total of 3 locations where colonoscopies were performed. R: To increase number of eligible patients.
	Appointment Made on Date (added)	Appointment Made on Date	D: Constrains the algorithm to search for appointments made within a specified date range. P: Set to 'equal to T-1' (or T-3 on Monday), so that only appointments made on the previous business day are captured. R: Eliminated multiple outputs of the same appointment. Ensured all newly made appointments would be captured.
		Appointment Cancel Date (added)	D: Constrains the algorithm to search for appointments cancelled within a specified date range. P: Set to 'equal to T-1' (or T-3 on Monday), so that only appointments cancelled on the previous business day are captured. R: Appointment Cancel Date restricted the FIT Daily Report to only capture cancellations that were made for appointments made and cancelled on T-1 (or T-3 on Monday), missing cancellations for appointments not made on that day. To capture all cancelled appointments a new report "FIT Daily Cancellations Report" was created to be run alongside the FIT Daily Report, where Appointment Cancel Date replaced Appointment Made on Date.
	Contact Type (added)		D: Type of contact patient had with hospital (115 types, such as 'Admission,' 'Documentation,' and 'Nursing Home Visit'). P: Set to include only 'Appointment' to imitate the function of Appointment Status. Epic IT staff recommended this criterion be added to make the algorithm run faster. R: The FIT Daily Report would have run correctly without this criterion but adding an additional restriction greatly reduced the population of patients to be filtered through the algorithm.
		Contact Type (revised)	P: Expanded to also include 'Hospital Encounter.' R: Some colonoscopy reschedules were not being tracked after cancellation. Scheduling staff at the gastroenterology clinic explained that Contact Type 'Appointment' captured only outpatient scheduling of the procedures. Expanding to include 'Hospital Encounter' captured the missing colonoscopies that were being scheduled as inpatient procedures.
	ICD-10 Codes (added)	ICD-10 Codes	D: Classification of medical diagnoses. P: Set to 'does not contain'. R: 175 values to exclude patients with specific conditions from study, such as colitis.
	Patient Type (added)	Patient Type	D: Special populations, e.g., prisoners, group home residents, student athletes, etc. P: Set to 'does not exist'. R: Excludes all patients with a Patient Type designation. Nursing home residents were removed after manual review as they did not have a Patient Type code.

*All Appointments Report did not include the Appointment Made on/Cancel Date criteria, and the Date Range was set as 'from T to M + 6.'.

T, Today; W, Week; M, Month; Y, Year; n/a, not available; IT, information technology; FIT, fecal immunochemical testing; GI, gastroenterology.

Table 2. List of Epic Report Display Fields

Display Field Name	Description	First Iteration	Second Iteration	Third Iteration
MRN	Patient medical record number	✓	✓	✓
Status	Scheduling status: scheduled, cancelled, no show, arrived, completed, etc.	✓	✓	✓
Date	Date of colonoscopy	✓	✓	✓
First Name	Patient's first name	✓	✓	✓
Preferred Name	Patient's preferred name			
Last Name	Patient's last name	✓	✓	✓
DOB	Date of birth	✓	✓	✓
Age	Age	✓	✓	✓
Sex	Sex	✓	✓	✓
Race	Race	✓	✓	✓
Hispanic Ethnicity	Yes/No answers	✓	✓	✓
Preferred Language	Preferred language			✓
Phone	Primary phone number	✓	✓	✓
Home	Home phone number		✓	✓
Cell	Cell phone number			✓
Patient Email Address	Email address			✓
Patient Address (Line 1)	Patient address (line 1)	✓*	✓	✓
Patient Address (Line 2)	Patient address (line 2)	✓*	✓	✓
City	City	✓*	✓	✓
State	State	✓*	✓	✓
ZIP Code	ZIP code	✓*	✓	
Department	Department at which the procedure will be performed		✓	✓
Type	Type of colonoscopy procedure	✓	✓	✓
Ordering Provider	Ordering provider	✓	✓	✓
Attending Provider	Attending provider	✓	✓	✓
Associated Diagnosis	Associated Dx in the order for colonoscopy		✓	✓
Indication	Physician's indication for the Dx in the order		✓	✓
Rectal Bleeding†	Patient had rectal bleeding in the past 60 days			✓
ASN†	Appointment serial number—A unique identifier for all appointments associated with a physician order. Used in combination with CSN to track appointment changes.			✓
CSN‡	Contact serial number—A unique identifier for a new appointment. Used in combination with ASN to track appointment changes.			✓

*Address was a single field in the first iteration. Subsequent iterations imported discrete fields.

†The ICD-10 codes associated with rectal bleeding (K62.5, K92.1, and R19.5) were found in the encounter diagnoses from encounters that occurred in the past 60 days and grouped together to create a single variable, Rectal Bleeding. These were distinct from the ICD-10 codes criterion, which was designed to search the problem list.

‡In Epic, a unique Contact Serial Number (CSN) is generated for each new appointment that is made for an Order. The first CSN to be generated is designated as the Appointment Serial Number (ASN), which serves as the reference number connecting all subsequent appointment changes that occur for that Order.

the appointment. A full list of Output variables and their definitions can be found in Table 2.

The Epic algorithm was developed and refined through 3 iterations to retrieve a list of patients who met the eligibility criteria. The first iteration was developed for a pilot for the BestFIT study; consecutive iterations improved on this baseline. An Epic FIT Daily Report was developed to be manually triggered every business day to capture all colonoscopy appointments made on the previous business day. This FIT Daily Report served as a prescreening and provided a list of potentially eligible participants who then underwent a final manual chart review using additional criteria the Epic report could not accurately capture. Manual review included criteria such as reason for colonoscopy and time since last colonoscopy. See Appendix A for an in-depth discussion of the manual review process.

As an accompaniment to the EHR algorithm, researchers with the help of another IT group developed a custom candidate and participant tracking application in File Maker Pro (Claris International Inc., Santa Clara, CA) to process the Output from the FIT Daily Report (referred to as Tracking Database). When Output from the FIT Daily Report was imported to the Tracking Database, it performed additional critical functions for enhancing subject recruitment such as marking candidate eligibility, prompting recruitment mailing to eligible patients when they fell within the appropriate recruitment time frame, and automatically updating the appointment information when appointment dates were changed. Without the Tracking Database, the volume of patient records generated on a daily basis would have resulted in a much more labor-intensive patient review and recruitment process as the scheduling status updates would have had to have been tracked manually in a spreadsheet before recruitment activities could resume.

When the EHR algorithm development was done, 3 distinct algorithms had been created in Epic for comprehensive coverage of the complex nature of tracking scheduling events throughout the duration of the 5-year study: a FIT Daily Report, a FIT Daily Cancellations Report, and an All Appointments Report. The FIT Daily Report and FIT Daily Cancellations were developed to be run every business day; the All Appointments Report captured all existing appointments for the upcoming 6 months and was run once at the start of the BestFIT study to provide a data bank to which the 2 Daily Reports would update. The FIT Daily Report, its criteria, definitions, and parameters are described in Table 1.

Data Analysis
Each criterion for the Epic algorithm was tested for accuracy through a validation process. The development of the FIT Daily Report went through 3 iterations with new iterations built when researchers identified mistakes with the existing version.

After the algorithm was built and run for the first iteration, the algorithm was run again subtracting each criterion one at a time to verify its accuracy. Second and third iterations were tested by adding new or revised criteria to the algorithm one criterion at a time and testing its accuracy. Within each criterion, parameters were set using relationship commands to specify whether values were to be included or excluded. Examples of relationship commands for inclusion included "equal to," "greater than or equal to," or "less than or equal to." For exclusion criteria, relationship commands included "does not contain" or "does not exist."

Results were verified using 2 methods. The first method was to calculate and review descriptive statistics to see whether the Output values fit the conditions set in the criteria (SPSS version 25, IBM, Armonk, NY). For example, Age ranged from 50 to 85 years of age when the Age criterion was included in the FIT Daily Report. However, when the Age criterion was removed from the algorithm, the range was 18 to 90 years of age, verifying that the Age criterion was working correctly. Verification using descriptive statistics was appropriate for all variables except ICD-10 codes. To test the accuracy of the ICD-10 codes criterion, problem lists were reviewed manually to verify that those who were ineligible were excluded. This was because the medical diagnosis problem list could have as little as zero or up to as many diagnoses (20 to 30) as required by a patient. In this case it was not pragmatic or possible to have the Output file list all the ICD-10 codes included in the problem list for each subject (contact the authors for a complete list of ICD-10 codes).

Before the start of the BestFIT study, the research team evaluated and verified the algorithm. The original runs (referred to as original runs) were conducted throughout this time period with 1 or 2 weeks spent on verifying each criterion. The 1 to 2 weeks spent on verification generated enough patients scheduled for a colonoscopy for a thorough validation. For this article, researchers recreated the steps that were taken for the original runs and retroactively ran the FIT Daily Reports, but in a much more condensed manner (referred to as retroactive runs). Instead of running the iterations consecutively with multiple days spent on each iteration, all iterations (first, second, and third) were run simultaneously. To ensure enough data were present to test that each of the criteria were working correctly, the retroactive runs were executed over the course of 3 days, referred to as days 1, 2, and 3, respectively (Table 3). Column numbers varied by day depending on the number of colonoscopies that had been scheduled in the gastrointestinal clinic; row numbers varied depending on the criteria being tested.

Results
Each of the 3 iterations successfully built on one another to identify the eligible patients for the

BestFIT study. The first iteration served as the baseline where researchers identified the need to remove prisoners and patients living in assisted care facilities and improve the algorithm for generating eligible subjects for the study. The second iteration allowed for more accurate targeting of the desired population by applying filters for ICD-10 codes and expanding Appointment Procedures and Encounter Departments. Most importantly, the Appointment Made on Date filter introduced in the second iteration made it so that no scheduling changes would fall through the cracks while reducing the number of patients that required daily manual review in the Tracking Database. Further improvements were made in the third iteration to prevent any loss in potential candidates for the BestFIT study by extending the Date Range, developing a second identical Report that replaced Appointment Made on Date with Appointment Cancelled Date, and adding the Output elements Appointment Serial Number (ASN) and Contact Serial Number (CSN) to enable the Tracking Database to automatically update appointment changes.

The variation in the numbers within each iteration (ie, column) denote the changes brought on by the exclusion or inclusion of criterion, but there were cases where the numbers showed no change. Day 2 had 2 instances where the number of patient records did not vary between criterion tests. One case was in the first iteration of day 2. The numbers were identical when Appointment Status and Age criteria were removed from the algorithm, respectively (n = 276). Researchers were able to verify that this was coincidental. The second case was when Contact Type was revised in the third iteration of day 2. In this case, researchers could only surmise there had been no appointments made under the "Hospital Encounter" code, as adding this to the Contact Type did not change the numbers on this day (n = 14). Due to this second issue, only days 1 and 3 provide sufficient data to verify all steps of the iterations. For simplicity, we only discuss the outcome for day 1 shown in Table 3.

First Iteration
The first iteration included 5 criteria: Age, Appointment Procedure, Appointment Status, Encounter

Table 3. Verification of Criteria from Three Distinct Retrospective Runs

Iteration	Criteria	n Day 1	Day 2	Day 3
First Iteration*	Age, Appointment Procedures, Appointment Status, and Encounter Department	219	225	174
	All Criteria minus Age	267	276	215
	All Criteria minus Appointment Procedures	771	807	782
	All Criteria minus Appointment Status‡	270	276	228
	All Criteria minus Encounter Department§	§	§	§
	All Criteria with substitute Encounter Department	0	0	0
Second Iteration†	Patient Type	261	266	218
	Appointment Procedures‖	415	420	431
	ICD-10 Codes (Using Boolean Operator OR)¶	*415*	*420*	*431*
	ICD-10 Codes (Using Boolean Operator AND)	336	331	348
	Encounter Department‖	657	647	703
	Contact Type	530	544	531
	Appointment Made on Date	15	14	32
Third Iteration†	Contact Type‖	21	14	39
	Date Range	38	30	61

*Compare against n in top row.

†Each criterion is added sequentially. Compare against n of the immediately preceding criterion.

‡The second iteration began with the removal of Appointment Procedures. Therefore, values in this row serve as the baseline for the second iteration.

§The FIT Daily Report timed out due to restrictions in place from Epic Reporting Workbench and did not yield results. In this institution's Epic system, reports must include at least the Providers or Encounter Department field or have a patient base selected to prevent timing out.

‖These criteria are revised from the previous iteration.

¶See Appendix B for further details.

Departments, and Date Range. The Output yielded a total of 219 patient records with appropriate values for each criterion. Four criteria were tested by removing one criterion at a time to observe the change it wrought in the Output. Date Range was not tested at this time. When the Age criterion was removed, the number increased from 219 to 267, and the age range widened to include ages 19 through 90 years, explaining the larger number of patients in the report with age excluded. When Appointment Status, which limited the search to "Scheduled" appointments, was eliminated from the algorithm, the number of patients increased from 219 to 270. The difference was verified to be due to the cancelled colonoscopies that were not filtered out. Removing the Appointment Procedures criterion resulted in 771 records in the Output, and the Output yielded 33 other procedures that took place in the designated Encounter Department.

Finally, removing the Encounter Departments criterion resulted in the FIT Daily Report timing out and yielding an error with no results. The Report was timing out because the absence of an Encounter Department made the report take up too many resources in the Epic system. To verify the Encounter Department criterion, researchers substituted an unrelated department in the field instead of removing the criterion altogether. When the Encounter Department was changed to a noncolonoscopy procedure unit, there were zero results, as expected.

Second Iteration

The second iteration tested the following changes: Patient type was added to the report criteria to exclude individuals residing in prisons; Appointment Procedures and Encounter Departments were expanded to include more potential candidates; ICD-10 codes were added to exclude ineligible patients; Appointment Made on Date was included to improve the definition of the parameters of the Report. Appointment Status was removed as it was not capturing all the changes that could occur with an appointment. In the second and third iterations, criteria were tested one at a time, cumulatively.

With Appointment Status removed, 270 records comprised the baseline for the second iteration. Adding Patient type removed 9 records that were classified as residing in a correctional facility, resulting in 261 records. Adding an additional 10 Appointment Procedures increased the number to

415 (Table 4). It was important to discern the type of procedure being conducted during the colonoscopy as some colonoscopies were for stool transplant or completed through a stoma, which were not appropriate for this study. Using the Boolean operator "OR" between each of the ICD-10 codes resulted in no change in numbers but replacing it with the correct operator "AND" successfully reduced the number to 336 (see Appendix B for more detail). Expanding the number of Encounter Departments to 3 departments increased the resulting list to 657; restricting the Contact Type to just Appointments reduced the number to 530. This reduction in numbers was surprising as we expected the number to stay the same with the addition of

Table 4. Colonoscopy Procedure Names and Codes Used for Scheduling

	Procedure Name [Code Number]
Appropriate for Use	IRL COLONOSCOPY PROCEDURE [325]
	ENDO COLON COMPLEX- RN SED [1687]
	ENDO COLONOSCOPY [1685]
	ENDO COLONOSCOPY COMPLEX [1191]
	ENDO COLONOSCOPY COMPLEX-ANES [1373]
	ENDO COLONOSCOPY- ANES [1686]
	ENDO COLONOSCOPY-RN SED [1190]
	ENDO/COLON COMBO [1775]
	ENDO/COLON COMBO - ANES [1777]
	ENDO/COLON COMBO - RN SED [1776]
	ENDO COLON [328]
Inappropriate for Use	ENDO CLN W STENT PLCMNT-RN SED [1689]
	ENDO COLON STENT PLCMNT FLUORO [1194]
	ENDO COLON STENT PLCMNT-ANES [1375]
	ENDO COLON STENT PLT- ANES ONLY [1481]
	ENDO COLON W STENT PLACEMENT [1193]
	ENDO COLON W STL TRANS- ANES [1376]
	ENDO COLON W STOOL TRANSPLANT [1196]
	ENDO COLON/ILEO STOMA- ANES [1374]
	ENDO COLON/ILEO STOMA- RN SED [1688]
	ENDO COLON/ILEO THROUGH STOMA [1192]

the "Appointment" Contact Type. Upon further examination, we learned another value, "Hospital Encounters," met our criteria and needed to be added, which was done during iteration 3.

Finally, introducing Appointment Made on Date made the biggest impact on reducing the number of patients to review. With the algorithm limiting the output to just the patients who had made an appointment on the prior business day, the number went down to 15 indicating that was the number of colonoscopy appointments made on one day.

Third Iteration

After learning from gastroenterology scheduling staff that some of the appointments not being captured were because of an overlooked Contact Type value, researchers ran a third iteration where Contact Type "Hospital Encounter" was added. This revision yielded an increase from 15 to 21 records. Extending the Date Range from 12 weeks to 4 years increased the final output number to 38 patient records and thus captured appointment scheduled, cancelled, and rescheduled.

In summary, the algorithm was refined through 3 iterations. This was done sometimes by adding additional criteria values that were missed in previous iterations that led to an increase in potentially eligible patients (e.g., second iteration of Appointment Procedures and Encounter Departments) while at other times it was done by filtering out unwanted characteristics (e.g., Patient Type), thereby increasing the sensitivity of the algorithm (Table 3). By the third and final iteration, the sensitivity and specificity of the algorithm had been maximized. A total of 14 reports were generated each day (not counting the first iteration Encounter Departments report that timed out), with the number of patients per report being listed in the rows. All patient records in every report were reviewed manually for accuracy of the individual criterion being tested.

Discussion

In our literature review, no studies were found where EHR reports were combined with patient appointment information and perfected to where reports were built specifically around appointment cancellations and reschedules to aid in study recruitment. This study is unique in that while administrative data and patient data are routinely used for research recruitment, adding the appointment date, and automatically tracking the cancellation and rescheduling of appointments is novel. Other researchers have attempted it but through manual reviews of the appointment updates, rather than through an algorithm including appointment rescheduling and cancelling.[4] Capturing and up-dating appointment cancellations and reschedules optimizes subject recruitment compared with recruitment based solely on scheduled appointments. The appointment date and cancelling or rescheduling of the appointment date were extremely important to include in the algorithm as the BestFIT study invited patients to participate in the study anywhere from 28 to 56 days out from the scheduled procedure. This time frame allowed for mailing of the informed consent, receipt of the signed consent, mailing of the occult blood tests, and receipt of the completed tests before the colonoscopy.

Having an accompanying Tracking Database was crucial to using the appointment change information. The Tracking Database processed the ASN and CSN to update changes to appointments so that the BestFIT study could dynamically adjust timeframes for subject contact. This ability to track the changes to appointments is especially useful for studies where the timing of the medical appointment is critical to the study, as it was in the BestFIT study, since FITs needed to be collected ahead of the colonoscopy prep and at most 4 months out from the colonoscopy. It also allows researchers to recruit more subjects in a shorter time period.

Researchers went through multiple cycles of communications with an expert IT application developer to create the queries for identifying potential participants for the study. This was time-consuming and, similar to what has been found by other researchers, the queries were prone to errors which were found in the validation process.[1] Generally, researchers will most likely know the demographic and diagnostic parameters of the data they seek from the EHR, but may be less familiar with the many administrative criteria, such as appointment scheduling, identifying the departments in which the procedures take place, and setting correct date ranges—all which need careful consideration to recruit the desired population. Such was the case for this study and errors generally occurred due to lack of knowledge of administrative criteria and how they were used in the hospital setting (e.g., Appointment Procedures, Encounter

Department, and Contact Type). For instance, if we had not pre-emptively restricted the parameters of Contact Type in the second iteration to include only "Appointments," the wider net would have captured the "Hospital Encounters." We blindly added Contact Type to imitate the function of Appointment Status that was being replaced, and the significance of "Hospital Encounters" was not apparent to us until we noticed that some rescheduled appointments were missing.

EHRs capture data on an open cohort of patients that have entered the hospital system. Part of the rigor of this study was the investigation of each data field to ascertain accuracy and comprehensively capture the data that was necessary for recruitment. Since EHR data are not standardized nor intuitively structured, we found, similar to other researchers, the need to work closely with IT staff to carefully review and validate the criteria chosen for algorithm development.[10] In fact, it was IT staff who discovered the ASN and CSN that allowed the tracking of appointment schedule status—researchers would not have known to look for such a variable. Future researchers may have a more streamlined experience developing their algorithms if they take time to consult and plan carefully with appropriate IT and administrative staff to gain more than a cursory understanding of criteria and their values.

Many studies suffer from difficulty with patient recruitment or recruitment delays.[4,11–13] Patient recruitment is essential for study success. Without it, studies can be delayed with the potential loss of funding.[14] The time and effort invested in the recruitment algorithm for the BestFIT study was time-efficient in that nearly all potentially eligible patients were invited to participate, as indicated by careful validation of the iterative runs. This due diligence enhanced the actual recruitment process by excluding patients who were ineligible or who had changed their colonoscopy appointment date to be beyond the desired time frame.

Intuitively, it would follow that a thorough algorithm would make manual review of each eligible record unnecessary. However, it was impossible to build the algorithm to meet all the eligibility specifications due to lack of uniformity, completeness, and accuracy of the patient care data captured in the EHR. For instance, there was lack of uniformity in the terminology used by providers in noting the type of colonoscopy, some patients did not have records of past colonoscopies, and some clinic notes did not reference the most recent colonoscopy reports.

An inexperienced programmer resulted in initial errors in excluding ICD-10 codes for ineligible patients. This resulted in us diligently checking each criterion.

Conclusion

Although labor intensive, the time and effort put into the development of an EHR algorithm proved successful for recruitment with little effort to run each day. Once the BestFIT study was underway, it saved researchers' time by excluding many ineligible individuals before the manual review. With the help of IT, we refined, tested, and validated the algorithm. By including the capture of cancelled and rescheduled colonoscopy appointments, we ensured we were not missing potential participants. Fortunately, time was available for this development as the research team found out that the BestFIT study was likely to be funded 8 months before the actual start date. Different skill sets of the research team (nurse, physician, Epic database expert, and software developers) facilitated the pursuit and success of this endeavor. This work expands on what has been done in the literature, demonstrating that adding cancellations and reschedules of colonoscopies will optimize the potential number of eligible patients for recruitment.

To see this article online, please go to: http://jabfm.org/content/34/1/49.full.

References

1. Xu J, Rasmussen LV, Shaw PL, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc 2015;22:1251–60.
2. Samuels MH, Schuff R, Beninato P, et al. Effectiveness and cost of recruiting healthy volunteers for clinical research studies using an electronic patient portal: a randomized study. J Clin Transl Sci 2017;1:366–72.
3. Kopcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014;16: e161.
4. Joseph CL, Ownby DR, Zoratti E, et al. Recruitment experience for a pragmatic randomized controlled trial: using EMR initiatives and minimizing research infrastructure. Clin Res Regul Aff 2016;33:25–32.

5. Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clinical Trials 2012;9:198–203.

6. Badurdeen DS, Umar NA, Begum R, et al. Timing of procedure and compliance with outpatient endoscopy among an underserved population in an inner-city tertiary institution. Ann Epidemiol 2012;22:531–5.

7. Jackson DS, Egbuonnu N, Umunakwe C, et al. Scheduled out-patient endoscopy and lack of compliance in a minority serving tertiary institution. Am J Med Sci 2012;344:194–8.

8. Turner BJ, Weiner M, Yang C, TenHave T. Predicting adherence to colonoscopy or flexible sigmoidoscopy on the basis of physician appointment-keeping behavior. Ann Intern Med 2004;140:528–32.

9. Kazarian ES, Carreira FS, Toribara NW, Denberg TD. Colonoscopy completion in a large safety net health care system. Clin Gastroenterol Hepatol 2008;6:438–42.

10. Sweet LE, Moulaison HL. Electronic health records data and metadata: challenges for big data in the United States. Big Data 2013;1:245–51.

11. Oakes JM, MacLehose RF, McDonald K, Harlow BL. Using administrative health care system records to recruit a community-based sample for population research. Ann Epidemiol 2015;25:526–31.

12. Weng C, Appelbaum P, Hripcsak G, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc 2012;19:684–7.

13. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. BMC Med Inform Decis Mak 2008;8:13.

14. Wacher NH, Reyes-Sanchez M, Vargas-Sanchez HR, et al. Stepwise strategies to successfully recruit diabetes patients in a large research study in Mexican population. Prim Care Diabetes 2017;11:297–304.

Appendix A: Manual Review

After running the algorithm, a manual chart review was necessary because data constraints in the EHR prevented the algorithm from filtering out all ineligible patients. Data constraints included notes written as unstructured text (ie, there were no defined fields that the algorithm could identify), scanned PDF or image files received from other health care facilities, and problem lists that were not up to date. Examples of manually reviewed items are described below.

Clinic notes sometimes contained unstructured text that would indicate a colonoscopy was being ordered for diagnostic reasons. Often, these indications were not noted in the indication field of the physician colonoscopy order. The date of the most recent procedure note had to be read to ensure that at least 34 months had passed since the last procedure. This had to be done manually because the past procedure notes were sometimes found in the media tab as scanned images if the colonoscopy had been done at a different facility. Anemia is an example of an exclusion criterion that has an ICD-10 code but had to be manually reviewed because this is a problem that often stays on the problem list even after it is resolved. Many patients with a diagnosis of anemia were eligible after a chart review. Mental health diagnoses were reviewed, and if severe, those individuals were not invited to participate. For example, one potential participant had a phobia about receiving mail, so they were excluded from invitation. Medical diagnoses and surgical procedure lists were reviewed in case a colectomy or other gastrointestinal surgeries had been performed.

Appendix B: Discrepancies and Oddities

Throughout the iterations, 3 discrepancies were noted in the outputs which at first glance led the researchers to think the criteria filters were not working. However, IT examined these and found the algorithm was in fact working as intended, and the unexpected values were an artifact of the way Epic functioned. For instance, in many of the iterations, patients aged 86 years would seem in the output. This was because the algorithm filtered patients based on their age at the time of the FIT Daily Report being run but the Output displayed the age of the patient at the time of the colonoscopy. Therefore, patients who turned 86 years after the FIT Daily Report was run but before the time of the scheduled colonoscopy were included in the Reports. These patients were excluded from the study via the Tracking Database, which removed patients under the age of 50 years or over the age of 85 years at the date of colonoscopy.

A second example is where the Encounter Department would show a department other than 1 of the 3 procedure units that the researchers had designated, such as "Anesthesia Satellite." This happened because some patient orders included multiple department designations, particularly if the patient were to have anesthesia services. In these cases, the output would sometimes display the anesthesia department information instead of the procedure unit name. For these cases, the researchers reviewed the orders to confirm that they did include the correct Encounter Department(s) specified in the criterion.

The third example was when the Output of Appointment Status showed values other than the expected "Scheduled." Appointment Status was removed as a filter criterion after the first iteration of the algorithm, but it remained as a field in the Output. Reviewing the Outputs in the second and third iterations showed cases with "Cancelled" and "Completed" values in addition to the expected "Scheduled." These unexpected values appeared because the FIT Daily Report was set up to report the status of all scheduling events that occurred on T-1. For example, in day 2, a patient scheduled, then cancelled, an appointment on T-1, and both events were reported in the Output. Shuffles (eg, procedure room changes or appointment time changes) are considered a scheduling event as well, and if a shuffle occurred on the day of the colonoscopy, then the completed status will seem for that case. All results in Table 3 include the cases with these discrepancies.

A notable oddity in setting up the algorithm was the Boolean operator used for the ICD-10 codes. The Boolean operator "OR" was appropriate for all other criterion logic (except for Age) but using this logic operator on this criterion did not exclude the ICD-10 codes as expected. The expectation was that the "OR" statement would allow the algorithm to seek out patients that had any 1 of the 175 unwanted ICD-10 codes (eg, A or B or C) and exclude them from the search. However, unlike the other criteria that functioned in this manner, the ICD-10 code criterion required the use of "AND," meaning that the algorithm was directing the FIT Daily Report to include all 175 ICD-10 codes in its search for patients to exclude. This was not consistent with the other criteria, nor was it intuitive, and required specialized knowledge on the part of the IT staff to recognize the problem.