Research and Applications

Development and validation of techniques for phenotyping ST-elevation myocardial infarction encounters from electronic health records

Sulaiman Somani[1], Stephen Yoffie[1], Shelly Teng[1], Shreyas Havaldar[1], Girish N. Nadkarni[1,2,3], Shan Zhao[1,4], and Benjamin S. Glicksberg[1,5]

[1]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, New York, USA, [2]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA, [3]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA, [4]Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA, and [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Sulaiman Somani, Stephen Yoffie, and Shelly Teng contributed equally to this work.
Corresponding Author: Benjamin Glicksberg, PhD, Icahn School of Medicine at Mount Sinai, 770 Lexington Ave., 14th Floor, New York, NY 10065, USA; benjamin.glicksberg@mssm.edu

Received 25 January 2021; Revised 7 June 2021; Editorial Decision 6 July 2021; Accepted 29 July 2021

ABSTRACT

Objectives: Classifying hospital admissions into various acute myocardial infarction phenotypes in electronic health records (EHRs) is a challenging task with strong research implications that remains unsolved. To our knowledge, this study is the first study to design and validate phenotyping algorithms using cardiac catheterizations to identify not only patients with a ST-elevation myocardial infarction (STEMI), but the specific encounter when it occurred.
Materials and Methods: We design and validate multi-modal algorithms to phenotype STEMI on a multicenter EHR containing 5.1 million patients and 115 million patient encounters by using discharge summaries, diagnosis codes, electrocardiography readings, and the presence of cardiac catheterizations on the encounter.
Results: We demonstrate that robustly phenotyping STEMIs by selecting discharge summaries containing "STEM" has the potential to capture the most number of STEMIs (positive predictive value [PPV] = 0.36, $N = 2110$), but that addition of a STEMI-related International Classification of Disease (ICD) code and cardiac catheterizations to these summaries yields the highest precision (PPV = 0.94, $N = 952$).
Discussion and Conclusion: In this study, we demonstrate that the incorporation of percutaneous coronary intervention increases the PPV for detecting STEMI-related patient encounters from the EHR.

Key words: electronic health records, phenotyping, myocardial infarction, cardiology, big data

INTRODUCTION

With the introduction of electronic health records (EHRs) into health systems, there has been a substantial increase in the accessibility of patient data for development of decision support systems and new avenues of clinical research.[1–5] However, EHRs are primarily architected for optimal clinical workflow and to organize documentation for billing, which does not seamlessly integrate with most research pipelines.[6] Disease cohorts are often primarily defined through International Classification of Disease (ICD) billing codes, and while relatively straightforward, often are not sufficiently accu-

LAY SUMMARY

Classifying acute myocardial infarction phenotypes in electronic health records remains an unsolved problem. Disease cohorts defined through International Classification of Disease (ICD) billing codes are often imprecise, while more complex rule-based systems which leverage mult-imodal data require expert clinical input and are difficult to implement at scale. However, filtering for the presence of certain downstream clinical actions can increase specificity without making large trade-offs in sensitivity. We designed and validated algorithms to phenotype ST-elevation myocardial infarction (STEMI) by using discharge summaries, diagnosis codes, electrocardiography readings, and the presence of cardiac catheterizations on the encounter. Our algorithms use cardiac catheterizations to identify not only patients with an STEMI, but the specific encounter when it occurred.

rate.[7,8] Pre-existing electronic phenotyping algorithms utilize various modalities of clinical data, such as ICD codes, hospital-specific diagnosis and procedure codes, patient notes, medications, laboratory values, imaging, and biopsy results,[1,7,9,10] to design rule-based systems. However, these algorithms need to be manually built and validated by experts; thus, extracting clinically relevant insights that are broadly applicable from these large repositories becomes tenuous.[4,11]

In particular, while cardiovascular medicine has benefited from EHR-based research,[12–15] significant challenges still remain in this domain.[15] Of note, phenotyping disease severity is paramount given the impact that different subtypes of a single disease can have on clinical management. Many cardiovascular disease states such as heart failure, atrial fibrillation, and acute coronary syndrome cover a spectrum of clinical diagnoses that are not limited to a single phenotype, which makes identifying them from EHRs particularly difficult.[15] For example, the classical diagnosis of a “heart attack” is actually composed of multiple phenotypes under the broad category of acute coronary syndromes (ACS): ST-elevation myocardial infarctions (STEMIs), non-ST-elevation myocardial infarctions (NSTEMIs), and unstable angina (UA). While each of these diseases fall under the broad category of ACS, they manifest different ways and generally require different treatment courses. Notably, STEMIs are diagnosed when elevated biomarkers of cardiac injury (troponins) are present in conjunction with symptoms of myocardial ischemia (ie, retrosternal chest pain) and/or ST-segment elevations greater than 1.5 mV on a patient’s electrocardiogram (ECG).[16] However, implementing this clinical rule broadly in EHRs is non-specific to STEMI patients, since nearly 47% of patients have normal troponin values on admission.[17] Symptomatology is further difficult to capture sensitively, since certain patient subgroups (eg, females, diabetics) can present atypically without chest pain. Furthermore, since an ECG is not as natively integrated into an EHR as other types of data (eg, patient notes, labs, ICD codes), past studies have failed to evaluate the performance of incorporating clinician-confirmed readings into their phenotyping algorithms. To our knowledge, current algorithms have achieved positive predictive values (PPVs) of no greater than 82%, indicating the lack of a gold standard to robustly identify STEMIs from EHRs at scale using commonly available modalities.[18] Previous studies have also identified patients with a history of a myocardial infarction, but some have failed to assess performance during unique hospital admissions or encounters when these STEMIs took place.[18]

In order to further refine the occurrence of a STEMI to a single patient encounter, which can enable building cohorts of patient hospitalizations from which diagnostic, therapeutic, and quality improvement analytics may be obtained, the downstream clinical actions carry valuable insight that can be leveraged either as solitary markers in EHRs for diagnosing STEMIs. Utilizing the clinical course of a patient engenders greater specificity in these pre-existing algorithms to increase their PPVs without making large trade-offs in sensitivity. In large hospital systems which have cardiac catheterization laboratories, the management of a STEMI is an emergent percutaneous coronary intervention, which includes cardiac catheterization of a patient to visualize the coronary arterial tree and perform an intervention (ie, angioplasty, stenting) to liberate the obstruction causing the STEMI. In this study, we take advantage of EHR data, specifically a cardiac catheterization procedure record in a patient’s hospital admission, to investigate the benefits it affords for identifying hospital admissions for STEMIs from a large, multicenter health system. Additionally, we focus on isolating by unique patient encounters (ie, hospital admissions) associated with a STEMI, not only patients who have ever had a STEMI on any or their first patient encounter.

MATERIALS AND METHODS

Data sources

This study was conducted at the Mount Sinai Health System (MSHS), which is a comprehensive, multicenter hospital system based in New York City. We utilized an EHR from the diverse patient population at MSHS which contains records for over 5.1 million individuals and 115 million encounters, between January 2000 and June 2020. These records contain a variety of patient data, such as patient demographics, diagnosis codes, encounter details, provider notes, laboratory data, and procedure codes. Each of these data elements are linked with a hospital encounter and unique patient ID; however, association to the appropriate hospital encounter during which the STEMI took place may be subject to errors and are addressed in details below. This study was approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai.

Identification of STEMI

Potential cases of STEMI were identified by querying the EHR database for all records until June 17, 2020 using broad criteria such as indications of a STEMI in an ECG, ICD codes, or discharge diagnosis (Figure 1). These criteria were selected based on consultation on STEMI identification criteria in the clinical setting with two clinicians (one board certified Internal Medicine chief resident and one Anesthesiology resident) and review of prior algorithms.[15] Although a STEMI can be a part of a patient’s past medical history, the utility of diagnosing it for the purposes of this study was to identify the specific patient encounter in which the STEMI took place, not whether a patient had ever previously been diagnosed with a STEMI. Therefore, encounters requiring multiple criteria to be met

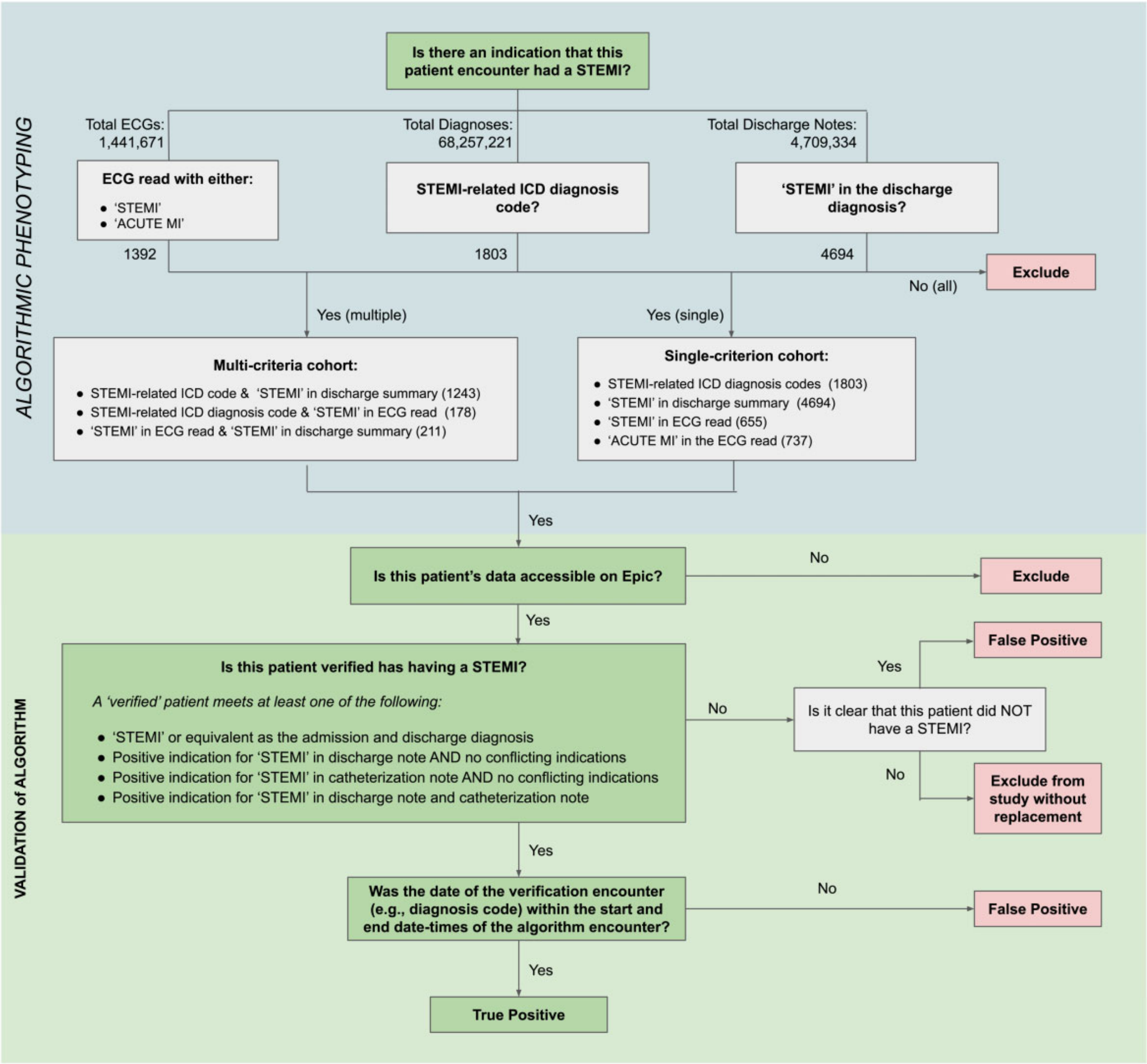
ALGORITHMIC PHENOTYPING
Is there an indication that this patient encounter had a STEMI?
Total ECGs: 1,441,671
Total Diagnoses: 68,257,221
Total Discharge Notes: 4,709,334
ECG read with either:
• 'STEMI'
• 'ACUTE MI'
STEMI-related ICD diagnosis code?
'STEMI' in the discharge diagnosis?
1392
1803
4694
Exclude
No (all)
Yes (multiple)
Yes (single)
Multi-criteria cohort:
• STEMI-related ICD code & 'STEMI' in discharge summary (1243)
• STEMI-related ICD diagnosis code & 'STEMI' in ECG read (178)
• 'STEMI' in ECG read & 'STEMI' in discharge summary (211)
Single-criterion cohort:
• STEMI-related ICD diagnosis codes (1803)
• 'STEMI' in discharge summary (4694)
• 'STEMI' in ECG read (655)
• 'ACUTE MI' in the ECG read (737)
Yes
VALIDATION of ALGORITHM
Is this patient's data accessible on Epic?
No
Exclude
Yes
Is this patient verified has having a STEMI?
A 'verified' patient meets at least one of the following:
• 'STEMI' or equivalent as the admission and discharge diagnosis
• Positive indication for 'STEMI' in discharge note AND no conflicting indications
• Positive indication for 'STEMI' in catheterization note AND no conflicting indications
• Positive indication for 'STEMI' in discharge note and catheterization note
No
Is it clear that this patient did NOT have a STEMI?
Yes
False Positive
No
Exclude from study without replacement
Yes
Was the date of the verification encounter (e.g., diagnosis code) within the start and end date-times of the algorithm encounter?
No
False Positive
Yes
True Positive

Figure 1. Phenotyping algorithm and validation procedure.

were merged based on the start date of the associated hospital encounter for that particular data element. For example, an ICD diagnosis code must have been assigned on the same encounter where a discharge summary containing the string "STEMI" was present, not on different encounters but for the same patient. Additional nuances of the date used for merging are discussed below for each criterion where appropriate. To simplify the analysis, we only assess for STEMIs that occurred during a formal hospital admission (eg, eliminating office visits where a STEMI could have occurred and was assigned but required subsequent hospitalization with another encounter as well) by filtering for all encounters with a discharge summary note type present.

Processing catheterizations

Catheterizations were retrieved from the EHR by querying for all encounters that were associated with two unique, database-specific procedure IDs. Manual inspection revealed that the associated Encounter ID (EID) of a catheterization was not always linked with the actual EID associated with the hospitalization where the potential STEMI took place. For this reason, the date of the catheterization was used for merging with other criteria by associating this catheterization to the in-hospital patient encounter that took place within 24 h of the catheterization date. Since catheterizations were merged with other criteria on a basis of time difference between the date of procedure and date of the concomitant criterion, we performed a sensitivity analysis to investigate the effect of increasing the time difference for filtering out non-qualifying encounters. These concomitant criterion included STEMI-associated ICD code, ECG read with "STEMI" keywords, and the presence of an associated hospital encounter. The number of unique patient encounters were calculated at varying absolute time differences of one, two, and four days for each secondary event.

Processing ICD codes

To identify STEMIs by ICD codes, the following ICD-9 and ICD-10 codes were used: 410, 410.21, 410.31, 410.41, 410.01, 410.11, 410.51, 410.61, 410.81, 410.91, I21, I21.11, I21.19, I21.21, I21.01, I21.02, I21.09, I21.29, and I21.3.[19,20] The EID associated with the ICD code assignment was also recorded, but was not always reliable. Since these ICD codes were part of the physician-recorded "Problem List," but not necessarily linked to the ICD codes assigned for an encounter after discharge for billing purposes, our ICD data tables contained codes that were assigned to patients for a principal STEMI-related encounter on a subsequent encounter (thus different EID than the STEMI-related encounter), thus prohibiting the use of the EID link between the ICD code and the relevant STEMI-related encounter. To limit this issue, all ICD codes were filtered to ensure that the start date of the encounter when the ICD code assigned was within 24 h of an inpatient hospital admission or emergency department visit that patient had in their medical record.

Processing discharge summaries

To identify STEMIs by provider-written notes, we filtered the EHR system by all discharge summaries in the EHR that included the following regular expression criteria: case-insensitive "STEMI" or case-insensitive "ST*elevat" or case-insensitive "ST*segment," where * is a wild-card character that can take on any value. Each of the expressions was used in every algorithm that included regular expressions, unless otherwise specified, to capture the greatest number of true positive patients. Heuristically, because discharge summaries were more rigorously linked to patient encounters than the ICD codes and catheterizations (above), the start date of the encounter associated with discharge summaries was set as the start date of the associated hospital encounter in the EHR.

Retrieval of electrocardiograph reads

Generally, ECGs contain basic patient record information, key demographics, vital signs, automated cardiac parameters (eg, heart rate, interval lengths), and automatically generated, but clinician confirmed, annotations of the ECG ("ECG reads"). To identify STEMIs by ECG reads, we filtered the ECG read by all case-insensitive mentions of the word "STEMI" in the ECG read. Empirically, we also noted that many patient encounters with STEMI were captured with less specific keywords. For this reason, we also investigated the presence of "ACUTE MI" in the ECG read as another criterion for evaluation. Only ECGs taken within 24 h of admission were retained for analysis to avoid capturing STEMIs as secondary complications in the hospital stay.

Data post-processing

All criteria were processed to include unique encounters, as defined by the Medical Record Number (MRN) and admission date. Additionally, several patients received multiple ECG readings within the same encounter and sometimes even the same day. If a patient encounter had at least one ECG read with "STEMI" in the text, the encounter was marked as a positive case and included only once in the filtered sample to then undergo the validation process as a true or false positive.

Algorithm validation

A randomized sample of approximately 50 unique encounters for 50 unique patients were retrieved for each criteria and manually investigated in the patient chart, using the following patient information: ECG reads, catheterization notes, and discharge summaries to confirm the presence or absence of a STEMI.[19,21] These criteria were selected through consultation with expert clinicians and review of prior algorithms. For criteria that included a positive case of catheterization in the patient encounter, we further evaluated a more exhaustive set of 150 records to more robustly estimate PPV for those criteria. In order to estimate the efficacy of sample size in determining PPV, we compared the PPV from a randomly selected first set of 50 cases for each set of criteria to the overall PPV of the full 150 cases. A complete overview of this analysis is included in the Supplementary Methods and Table S1. During validation, records that were inaccessible due to access privilege issues or those without a clear diagnosis or absence of STEMI were excluded from the batch of 50 without replacement from the original pool. To eliminate variation in the interpretation of chart evaluation, a single reviewer assessed each case for a clearly identified diagnosis, as indicated within the catheterization notes, discharge summaries, and evaluation of the ECG. Cases that resulted in a diagnosis other than STEMI were labeled as a false positive and diagnoses that were significantly ambiguous (ie, discharge notes reference both NSTEMI and STEMI throughout course of admission) were excluded from analysis without replacement. Full details on the clinical protocol and examples of true-positive and false-positive cases are presented in the Supplementary Methods and Figures S1–S6.

Statistical analyses

PPVs were calculated for all criteria as the proportion of confirmed STEMI cases (true positives) relative to all potential STEMI cases (true positives and false positives). PPVs were generated for the following filtering criteria: STEMI-related ICD diagnosis codes, "STEMI" in discharge summary, "STEMI" in ECG read, "ACUTE MI" in the ECG read; STEMI-related ICD diagnosis code AND "STEMI" in discharge summary; STEMI-related ICD diagnosis code AND "STEMI" in ECG read; and "STEMI" in ECG read AND "STEMI" in discharge summary. Additionally, PPVs were calculated for all mentioned criteria when merged with catheterizations as well. To assess for how many potential STEMIs each criteria may have been captured if applied across the entire EHR rather than just in the validated sample, a theoretical number of STEMIs captured by the criteria was extrapolated from multiplying the estimated PPV by the number of total encounters across the entire EHR for that criterion.

RESULTS

A total of 315 402 466 procedures, 68 257 221 patient diagnoses, 4 709 334 discharge summaries, and 1 441 671 ECGs from a total of 4 913 952 patients were available for query from the EHR database. A total of 54 750 catheterizations associated with a hospital encounter; 737 and 655 encounters with ECG reads with the appropriate "ACUTE MI" and "STEMI" keyword, respectively; 4677 encounters with discharge summaries containing the appropriate STEMI keywords; and 1803 encounters with a STEMI-associated ICD diagnosis codes resulted from our query and used for downstream analysis.

Impact of catheterization window time

In Figure 2, the number of patient encounters totaled 54 750 for 1 day; 58 607 for 2 days; and 62 807 for 4 days when the time difference was varied between catheterization date and hospital admission date for the associated encounter. For catheterization date and ICD code assignment date, the number of patient encounters were 1469 for a time difference of 1 day; 1532 for 2 days; and 1609 for 4 days. Finally, the number of encounters resulted for catheterization date and ECG read date were 437 for 1 day; 481 for 2 days; and 520 for 4 days.

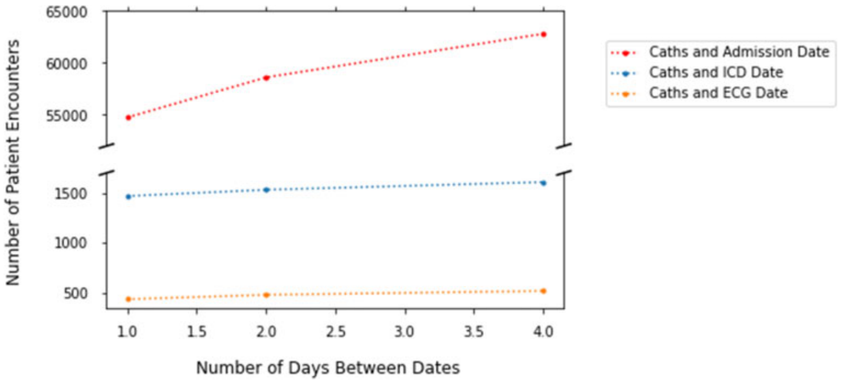

Figure 2. Effect of time (days) between catheterizations and secondary criteria on number of captured encounters.

Evaluation of independent criteria

Table 1 showcases the results of the first stage of manual validation for individual phenotyping criteria, as well as the increase in PPV afforded by inclusion of the presence of a cardiac catheterization on the encounter. The PPV for encounters associated with an ICD-10 diagnosis code for STEMI was 0.43, which captured 775 STEMIs in the EHR. Augmentation of this criteria with catheterizations increased the PPV to 0.76, which captured 1146 STEMIs despite lowering the total number of resulting records to 1469 from 1803. From the 4677 discharge summaries containing the appropriate search terms for a STEMI, a PPV of 0.36 was found on validation, which captured 2110 STEMIs. By the addition of a catheterization on the encounter, the PPV jumped to 0.73, with a total of 2285 encounters and 1168 STEMIs captured. Next, 737 ECGs were captured by a supported diagnosis of “ACUTE MI” on the ECG read, with a PPV of 0.33 that improved to 0.44 with the addition of the cardiac catheterization requirement to the encounter. Finally, 655 ECGs were captured by a supposed diagnosis of “STEMI” on the ECG read, with a PPV of 0.43 that resulted in 229 STEMIs being captured. Unlike other criteria, the addition of the catheterization constraint decreased the PPV to 0.36 and the number of STEMIs captured to 157, while also lowering the number of resulting records to 437.

Combining multiple criteria

We also assessed the ability to phenotype STEMIs by combining multiple criteria (Table 1). In addition, Figure 3 demonstrates the effect of merging these criteria on the total number of unique patient encounters for all combinations within the set, with and without catheterizations for each criteria. Filtering across the EHR for all records with an ICD code of STEMI and discharge summary returned 1243 encounters, with a PPV of 0.86 indicating that 1078 of them were STEMIs. Addition of the catheterization constraint increased the PPV to 0.94, but decreased the number of returning results to 1007 encounters with 952 potentially confirmed for STEMI based on this PPV. Next, the indicator of a STEMI in an ECG read was combined with STEMI ICD codes and found to have a PPV of 0.91 and a total of 163 potential STEMIs from 178 total resulting encounters. The addition of the catheterization constraint lowered the PPV to 0.86 and reduced the number of potential STEMI cases from ECG reads to 144 out of a total 169 resulting records. Finally, permutation of the above criteria to revolve encounters with a STEMI in the ECG read and STEMI in the discharge summary yielded 211 records, with a PPV of 0.75 and potential of 188 STEMI encounters. Augmentation with the catheterization constraint increased the PPV to 0.88, but with a lower total resulting encounters at 172 with only 155 of those as potential STEMIs. Though not incorporated as a potential phenotyping algorithm, the combination of all three criteria—ICD codes, discharge summaries, and “STEMI” ECG reads—yielded a total of 146 unique patient encounters, with only 137 resulting after the triad was combined with catheterizations.

DISCUSSION

This study aims to create and assess techniques to phenotype STEMI-related encounters from a multicenter EHR by examining the contribution of admission-specific procedures, ICD diagnosis codes, ECG reads, and discharge summaries in capturing STEMIs. Furthermore, we novely considered the effect of using catheterization in a patient encounter to augment performance. These criteria are nonspecific to MSHS and are generally applied at many large institutions, thus allowing for the translation of this algorithm across health systems. While the addition of catheterizations did decrease the number of unique patient encounters for a given set of criteria, the scale of decrease was not substantial (Figure 3). PPVs across all criteria without catheterizations ranged from 0.33 to 0.91, which rose to 0.36 to 0.94 with the incorporation of catheterizations. The magnitude of STEMIs captured from these algorithms is consistent with epidemiological data for the occurrence of STEMIs for the Northeastern United States.[22] To our knowledge, this is the first study to investigate the role of cardiac catheterizations, which is part of the first-line treatment for many patients with a STEMI, to increase the PPV of pre-existing criteria in determining hospital admissions associated with a STEMI.

Each individual criterion (eg, ICD codes, discharge summaries, and ECG reads) improved its PPV in characterizing STEMIs when requiring that the patient had a cardiac catheterization on that encounter. There was a significant improvement, for example, with ICD codes plus catheterization, in both PPV and number of STEMIs captured, despite a loss in the total number of resulting encounters. The same trend did not hold for the presence of STEMI in discharge summaries. Despite an increase in PPV, the total number of STEMI encounters captured was nearly 19% less than without incorporating catheterizations in the filtering criteria. This may likely be because the date of the catheterization was around the start of the hospital encounter, whereas discharge summaries may capture STEMIs that developed as complications during the inpatient stay or from various procedures. For the presence of “STEMI” in the ECG reads, the increase in total number of STEMIs captured, while mar-

Table 1. Performance of each phenotyping algorithm, as measured by PPV and the resulting total number of potential STEMIs captured in the EHR

	Catheterization	Number of encounters validated	Number of encounters excluded from validation	Total resulting encounters	PPV	Potential STEMIs captured	Potential false positives
ICD-10 Codes	No	49	1	1803	0.43	775	1028
	Yes	136	14	1456	0.76	1146	310
“STEMI” in Discharge Summary	No	47	3	4694	0.36	2110	2584
	Yes	143	7	2200	0.73	1668	532
ICD Codes + Discharge Summary	No	49	1	1243	0.86	1078	165
	Yes	145	5	1007	0.94	952	55
“ACUTE MI” in ECG Read	No	48	2	737	0.33	243	494
	Yes	143	7	458	0.44	179	279
“STEMI” in ECG Read	No	49	1	655	0.43	229	426
	Yes	137	13	437	0.36	157	280
“STEMI” in ECG Read + ICD Codes	No	45	5	178	0.91	163	15
	Yes	140	10	169	0.86	144	25
“STEMI” in ECG Read + Discharge Summary	No	48	2	211	0.75	188	23
	Yes	136	14	172	0.88	155	17

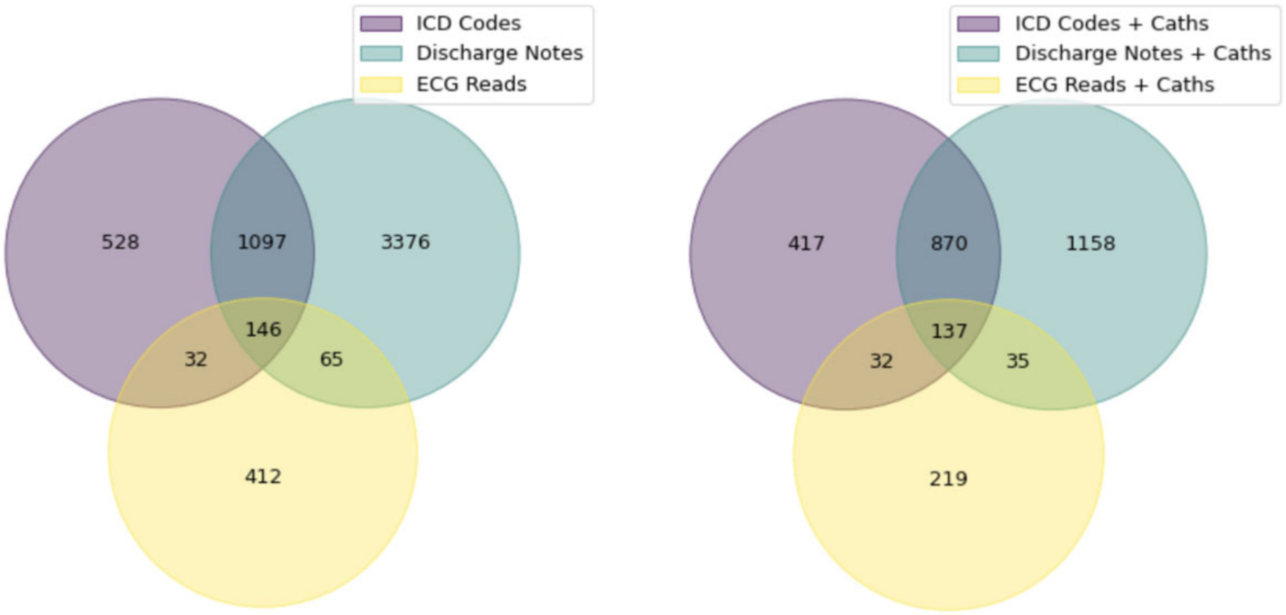

Figure 3. Venn diagrams demonstrating the overlap in encounters for each criteria both without (left) and with (right) associated catheterizations.

ginal, held when augmented with the requirement of a concomitant catheterization on that encounter. Additionally, we use the change in PPV (ΔPPV) to assess the marginal contribution of each criterion in our final algorithm and as a proxy for the importance of each criterion. We find that the greatest change in PPV occurs from incorporation of STEMI-related ICD-10 codes (ΔPPV 0.21), followed by STEMI in the discharge summary (ΔPPV 0.18) and catheterizations on the encounter (ΔPPV 0.08).

We also investigated the false positives for each of the criteria. Most patients that had a false positive encounter for a STEMI when filtered by STEMI-related ICD codes had so because the ICD code itself was assigned not on the encounter when the STEMI actually took place, but rather on a follow-up visit for the STEMI after the acute recovery of that event. The overwhelmingly predominant source of false negatives in the discharge summaries were patients who had a history of a STEMI. With the addition of catheterizations, inconsistent documentation (ie, record of a STEMI in certain parts of the encounter vs. NSTEMI in others) became the new and most predominant source of false positives.

Finally, the source of most false positives in ECG reads came from a confirmed reading indicating a STEMI had taken place, but the ECG criteria (>1.5 mm elevation in ST-segment in females, >2 mm elevation in ST-segment in males) for diagnosis had not been met. This is a reasonable finding, given that automatic reads from ECG systems must be extremely sensitive to STEMIs, since the consequences of missing a STEMI in a patient would be dire; therefore, these systems are calibrated to operate with increased sensitivity at the cost of losing specificity.[23] From investigating ECGs in other criteria during chart review, it is worth noting that many STEMIs were described on the ECG to have an "ST-segment abnormality" or "Consider wall ischemia" as key terms. Incorporation of these highly non-specific terms could increase the sensitivity of capturing STEMIs in ECGs, but at the risk that expanding the boundary to a broader term like "ischemia" or "acute myocardial infarction" would inadvertently capture NSTEMIs and unstable angina, the prevalence of which is much higher and the management (percutaneous coronary intervention) may be the same at academic centers like ours with large cardiac catheterization capabilities.

By PPV, the best performing query criteria was the combination of encounters with STEMI-related ICD codes, the presence of STEMI keyword in the discharge summary, and a cardiac catheterization on the encounter. Notably, without the presence of a cardiac catheterization, the predominant source of false positive cases came from patients who had a past history of a STEMI (a problem common to both of these filtering criteria), which suggests that the addition of a peri-encounter cardiac catheterization offers a complementary way of narrowing down likely encounters where an actual STEMI took place. While these results are promising and superior to those achieved in other studies in the United States by magnitude of potential STEMIs captured,[18] there is also a strong opportunity to improve the free-text search result by incorporating more formal natural language processing techniques to catch spelling errors, negations, medical colloquialisms, and modifier terms (such as "history of") to create more sophisticated algorithms for parsing for the appropriate note. We further note that algorithms in European systems using ICD codes achieve remarkable performance in phenotyping STEMI (PPV 0.96–1.00),[19] these algorithms fail to generalize well to our healthcare system (ICD-10 PPV 0.43),[24] likely from international differences between health and healthcare systems.[25–27]

There are several limitations to this study. Given that access to ICD codes at the hospital financial level was not available, the ICD codes used in this study were those that were captured in the patient chart by manual entry from physicians, which may not be as well captured as the nosologists working in billing departments who work full-time to this cause.[28,29] Additionally, augmenting the presence of a cardiac catheterization to improve the PPV is biased toward large medical centers with vast cardiac catheterization capabilities. The application of this algorithm in a rural setting may not be as well-captured, where the patient may present to a community-based hospital and be transferred to a larger hospital for the interventional management of their STEMI. Additionally, there exists no "gold standard" for validating the performance of these criteria beyond PPV, since other classification metrics (eg, sensitivity, specificity) require true- and false-negative counts that would be extremely laborious to ascertain. Validating on a small sample risks introducing sampling bias in the results of this study. Institutional practices that affect patient records leading to incomplete cases, patient transfers to outside hospitals, and provider bias may impede generalizability to other health systems. While we choose features in our algorithm that are clinically prevalent and oriented toward guideline-directed management of STEMIs at hospital centers with catheterization capabilities and test its performance in a large, urban, five-hospital health system, this algorithm has not been validated at an external institution to confirm its external validity. Finally, this study is biased toward hospital admissions with principal, admission diagnoses of STEMI given that the merging operations (namely, the association of a cardiac catheterization to an encounter) take place on the encounter admission date, which prohibits capturing STEMIs that occur as complications of an inpatient hospital stay.

CONCLUSION

Currently, no gold standard exists to identify specific hospital admissions with STEMIs from EHRs, particularly when attempting to distinguish between other acute coronary syndromes and other STEMI mimickers. In this study, we explore how incorporation of treatment strategy, namely a percutaneous coronary intervention (left heart cardiac catheterization), can increase the PPV of capturing STEMIs. While the strongest performing criteria for capturing STEMIs was the presence of "STEMI" in a discharge summary, its remarkably low PPV (0.36) gives way to favor augmentation of that filtering criteria with the presence of a cardiac catheterization, which achieves a PPV of 0.73. Additionally, the combination of STEMI-related ICD codes, the presence of STEMI keyword in the discharge summary, and a cardiac catheterization on the encounter yielded the highest PPV, further supporting the role of cardiac catheterization as a viable adjunct for phenotyping STEMIs. While more extensive and sophisticated algorithms can be designed to better parse out free text documents such as ECG reads and discharge summaries, the addition of a cardiac catheterization provides a unique type of informative content that can improve the specificity of an algorithm by removing false positives that may otherwise not be identifiable by other analytical means intrinsic to the original filtering criteria.

FUNDING

This study was supported by the National Center for Advancing Translational Sciences, Natonal Institutes of Health (U54 TR001433-05). GN is supported by R56-DK126930, R01DK127139, U01HG007278, U01HG009610, and U01DK116100.

AUTHOR CONTRIBUTIONS

SS, SZ, and BG conceived the idea. SS, SY, and ST performed data curation and analysis. SY performed manual chart validation. SS, SY, and ST drafted the manuscript. All authors contributed to the design, refinement of the framework, and final draft of the manuscript.

ETHICS STATEMENT

This study was approved by the Mount Sinai Institutional Review Board (IRB) 19-00951.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank the Mount Sinai Data Warehouse team for providing data.

CONFLICT OF INTEREST STATEMENT

SS is a co-founder and equity owner of Monogram Orthopedics. GN reported being a scientific co-founder, consultant, advisory board member, and equity owner of Renalytix AI, is a scientific co-founder and equity holder for Pensieve Health, being a consultant for Variant Bio and receiving grants from Goldfinch Bio and receiving personal fees from Renalytix AI, BioVie, Reata, AstraZeneca, and GLG Consulting. The other authors do not report any conflicts of interest.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly to maintain patient privacy, in accordance with the HIPAA Security Rule.

REFERENCES

1. Kim E, Rubinstein SM, Nead KT, *et al.* The evolving use of electronic health records (EHR) for research. *Semin Radiat Oncol* 2019; 29 (4): 354–61.
2. Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform* 2018; 22 (5): 1589–604.
3. Landi I, Glicksberg BS, Lee H-C, *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med* 2020; 3: 96.
4. Rasmussen LV, Thompson WK, Pacheco JA, *et al.* Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *J Biomed Inform* 2014; 51: 280–6.
5. Glicksberg BS, Johnson KW, Dudley JT. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Hum Mol Genet* 2018; 27 (R1): R56–62.
6. Chi J, Bentley J, Kugler J, *et al.* How are medical students using the Electronic Health Record (EHR)?: An analysis of EHR use on an inpatient medicine rotation. *PLoS One* 2019; 14 (8): e0221300.
7. Wei W-Q, Teixeira PL, Mo H, *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016; 23 (e1): e20–7.
8. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015; 7 (1): 41.
9. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019; 7 (2): e12239.
10. Sharma DK, Solbrig HR, Prud'hommeaux E, Pathak J, Jiang G. Standardized representation of clinical study data dictionaries with CIMI archetypes. *AMIA Annu Symp Proc* 2016; 2016: 1119–28.
11. Kagawa R, Shinohara E, Imai T, *et al.* Bias of inaccurate disease mentions in electronic health record-based phenotyping. *Int J Med Inform* 2019; 124: 90–6.
12. Abrahão MTF, Nobre MRC, Gutierrez MA. A method for cohort selection of cardiovascular disease records from an electronic health record system. *Int J Med Inform* 2017; 102: 138–49.
13. Selvaraj S, Fonarow GC, Sheng S, *et al.* Association of electronic health record use with quality of care and outcomes in heart failure: an analysis of get with the guidelines-heart failure. *J Am Heart Assoc* 2018; 7 (7): e008158.
14. Hulme OL, Khurshid S, Weng L-C, *et al.* Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC Clin Electrophysiol* 2019; 5 (11): 1331–41.
15. Hemingway H, Asselbergs FW, Danesh J, *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur Heart J* 2018; 39 (16): 1481–95.
16. Kala P, Mates M, Želízko M, *et al.* 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: summary of the document prepared by the Czech Society of Cardiology. *Cor et Vasa* 2017; 59 (6): e613–44. doi:10.1016/j.crvasa.2017.10.008
17. Wanamaker BL, Seth MM, Sukul D, *et al.* Relationship between troponin on presentation and in-hospital mortality in patients with ST-segment-elevation myocardial infarction undergoing primary percutaneous coronary intervention. *J Am Heart Assoc* 2019; 8 (19): e013551.
18. Rubbo B, Fitzpatrick NK, Denaxas S, *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: a systematic review and recommendations. *Int J Cardiol* 2015; 187: 705–11.
19. Coloma PM, Valkhoff VE, Mazzaglia G, *et al.* Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. *BMJ Open* 2013; 3 (6): e002862.
20. Patel AB, Quan H, Welsh RC, *et al.* Validity and utility of ICD-10 administrative health data for identifying ST- and non-ST-elevation myocardial infarction based on physician chart review. *CMAJ Open* 2015; 3 (4): E413–8.
21. Yeh RW, Sidney S, Chandra M, *et al.* Population trends in the incidence and outcomes of acute myocardial infarction. *N Engl J Med* 2010; 362 (23): 2155–65.
22. Ward MJ, Kripalani S, Zhu Y, *et al.* Incidence of emergency department visits for ST-elevation myocardial infarction in a recent six-year period in the United States. *Am J Cardiol* 2015; 115 (2): 167–70.
23. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. *J Am Coll Cardiol* 2017; 70 (9): 1183–92.
24. Steinberg BA, French WJ, Peterson E, *et al.* Is coding for myocardial infarction more accurate now that coding descriptions have been clarified to distinguish ST-elevation myocardial infarction from non-ST elevation myocardial infarction? *Am J Cardiol* 2008; 102 (5): 513–7.
25. Bekelman JE, Halpern SD, Blankart CR, *et al.* Comparison of site of death, health care utilization, and hospital expenditures for patients dying with cancer in 7 developed countries. *JAMA* 2016; 315 (3): 272–83.
26. Ridic G, Gleason S, Ridic O. Comparisons of health care systems in the United States, Germany and Canada. *Mater Sociomed* 2012; 24 (2): 112–20.
27. Michaud P-C, Goldman D, Lakdawalla D, *et al.* Differences in health between Americans and Western Europeans: effects on longevity and public finance. *Soc Sci Med* 2011; 73 (2): 254–63.
28. Yu AYX, Quan H, McRae AD, *et al.* A cohort study on physician documentation and the accuracy of administrative data coding to improve passive surveillance of transient ischaemic attacks. *BMJ Open* 2017; 7 (6): e015234.
29. McCarthy C, Murphy S, Cohen JA, *et al.* Misclassification of myocardial injury as myocardial infarction: implications for assessing outcomes in value-based programs. *JAMA Cardiol* 2019; 4 (5): 460–4.