

Validating Acute Coronary Syndrome ICD Codes as Training Labels for Clinical AI

Daniel K. Amponsah

`emergency-care-hsr-dcc@stanford.edu`

Stanford Cardiovascular Institute

Gabrielle Bunney

Stanford University

Kate Miller

Stanford University

Ariadna Garcia

Stanford University

Anna Graber-Naidich

Stanford University

Keejeong Ryu

Stanford University

Himani Darji

Stanford University

Marilyn Ndukwe

Stanford University School of Medicine

Valerie J.F. McDonald

Stanford University School of Medicine

William M. Mehring

Stanford University

Jada Henderson

Stanford University

David Lei

Stanford University

Marium Rizvi

Stanford University

Adenine Cembellin-Kao

Stanford University

Sylvia D. Kreibig

Stanford University

Celina M. Yong

Stanford Cardiovascular Institute

William F. Fearon

Stanford Cardiovascular Institute

Maame Yaa A.B. Yiadom

Stanford University

Article

Keywords:

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11156040/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence models may improve emergency department (ED) screening for acute coronary syndrome (ACS), but such models often rely on International Classification of Diseases (ICD) codes as training labels, whose accuracy in contemporary ED populations is uncertain. In this 10-year retrospective study of 619,599 adult ED encounters, we compared ICD-coded ACS against a chart-reviewed reference standard defined as the treating team's documented diagnosis. Sensitivity, specificity, and positive predictive value were 94.4%, 99.7%, and 72.4%, respectively. Disagreement occurred in 0.31% of encounters. Among false positives (n=1,647), the leading reason was ACS suspected but later ruled out (80.3%); among false negatives (n=257), coder omission of documented ACS (86.0%). ACS ICD codes are suitable training labels for screening models where missing ACS is more consequential than over-identification, and the observed error patterns are systematic and interpretable rather than random.

INTRODUCTION

Artificial intelligence (AI) models are increasingly being developed to support clinical decision-making in the emergency department (ED), where rapid identification of high-risk conditions such as Acute coronary syndrome (ACS) can meaningfully change outcomes¹⁻³. The growing availability of structured electronic health record (EHR) data has made large-scale model development feasible, but the quality of any such model is bound by the quality of the labels used to train it which is a well-established constraint in supervised machine learning.⁴

International Classification of Diseases (ICD) codes are billing-mandated, structured, and routinely collected across health systems making them an attractive labeling source for large retrospective datasets. However, their primary purpose is administrative rather than clinical as they are assigned after the encounter, often by coders interpreting the clinical team's documentation^{5,6}. Prior work has shown that ICD-based ascertainment can vary in its alignment with the clinical record^{7,8}. This variability may be acceptable in some cases. When ICD codes are used as outcome labels to train clinical prediction models, however, an imperfect proxy for the clinical diagnoses can propagate systematic error into model behavior.⁹⁻¹¹

ACS is a clinically important example of this problem. ACS is a time-sensitive condition, especially its most lethal subtype, ST-segment elevation myocardial infarction (STEMI), which requires a 12-lead electrocardiogram (ECG) within 10 minutes of first medical contact¹². Treatment delays in ACS are associated with increased morbidity and mortality¹³⁻¹⁵. Immediate ECG evaluation of every arriving ED patient, however, is not operationally practical. This limitation has generated interest in AI-based tools that could identify patients at higher likelihood of ACS using information available upon arrival to the ED^{16,17}. In this screening context, failing to flag a patient with ACS is more consequential than obtaining

an ECG in a patient ultimately found not to have ACS. This places particular weight on the sensitivity of the labels used in model training, as opposed to their specificity.

ACS ICD codes are not unique to AI; they are also used throughout cardiovascular outcomes research, quality measurement, and registry-based surveillance.¹⁸ Despite the practical advantages of ICD-based labeling, the accuracy of ACS ICD codes has not been well characterized in a contemporary adult ED population. Prior validation work has focused largely on inpatient registries and administrative claims databases, often outside of the United States, leaving the ED population, where ACS screening decisions are actually made, comparatively unexamined.^{18,19} Moreover, little is known about why ICD-based labels and clinical diagnoses differ, which is essential for understanding how labeling error affects model performance and anticipated clinical consequences upon implementation.

We conducted a retrospective study of adult patients presenting to the ED at a tertiary academic medical center to evaluate agreement between the final hospital ICD-coded diagnosis of ACS and the final diagnosis documented by the treating team. We also performed a secondary analysis of discrepant cases to characterize sources of disagreement between ICD-coded diagnosis and the treating team’s final diagnosis. Beyond informing AI development, these findings bear on the broader use of ACS ICD codes as outcome definitions in cardiovascular research.

RESULTS

Study Cohort

Between January 1, 2015 and December 31, 2024, 647,499 adult ED encounters were identified. After exclusion of trauma and stroke activations (n = 27,900), 619,599 encounters were included in the final analysis. Of these, 6,471 encounters met at least one criterion for chart review: (1) presence of an ACS ICD code (n = 5,961), or (2) abnormal troponin with coronary artery disease (CAD) ICD code (n = 510). The remaining 613,128 encounters were treated as non-ACS based on the absence of both an ACS ICD code and the enrichment criteria described in the Methods (see “Selection of Encounters for Chart Review”; Fig. 1).

Cohort Characteristics

Chart-reviewed patients were older than the broader ED population, with a mean age of 69.3 years (SD 15.5) compared with 50.9 years (SD 20.8) among all other encounters. They were also more likely to be male (59.2% vs 44.5%). Medicare was the most common insurance among reviewed patients (59.7%), compared with 30.6% in the remainder of the ED population, consistent with the older age of the reviewed cohort. The racial distribution was broadly similar between groups, though reviewed patients were more often non-Hispanic (83.3% vs 70.9%). Table 1 reports additional demographic details.

Table 1
Demographic characteristics of patients in encounters identified for review versus all other encounters

Characteristic	Chart reviewed patients N = 6,471¹	All other patients N = 613,128¹
Age		
Mean (SD)	69.3 (15.5)	50.9 (20.8)
Median (Q1, Q3)	70.0 (59.0, 81.0)	50.0 (33.0, 67.0)
Min, Max	19, 106	18, 123
Sex		
Male	3,828 (59.2%)	272,905 (44.5%)
Female	2,642 (40.8%)	339,820 (55.4%)
Non-Binary	0 (0.0%)	2 (0.0%)
Unknown	1 (0.0%)	401 (0.1%)
Race		
White	2,976 (46.0%)	252,940 (41.3%)
Black or African American	480 (7.4%)	45,345 (7.4%)
Asian	1,292 (20.0%)	92,638 (15.1%)
Other	1,409 (21.8%)	195,783 (31.9%)
Unknown/Decline to State	314 (4.9%)	26,422 (4.3%)
Ethnicity		
Hispanic/Latino	1,013 (15.7%)	169,915 (27.7%)
Non-Hispanic/Non-Latino	5,391 (83.3%)	434,760 (70.9%)
Unknown/Decline to State	67 (1.0%)	8,453 (1.4%)
Insurance		
Private insurance	1,402 (21.7%)	210,524 (34.3%)
Medicare	3,865 (59.7%)	187,827 (30.6%)
Medicaid	1,076 (16.6%)	179,239 (29.2%)
Other	68 (1.1%)	15,812 (2.6%)
¹ n (%) unless otherwise specified		

Characteristic	Chart reviewed patients	All other patients
	N = 6,471¹	N = 613,128¹
Age		
Self-Pay/No Insurance	60 (0.9%)	19,726 (3.2%)
Language		
English	5,106 (78.9%)	498,884 (81.4%)
Spanish	576 (8.9%)	75,717 (12.3%)
Other	782 (12.1%)	36,909 (6.0%)
Unknown	7 (0.1%)	1,618 (0.3%)
Hypertension	48 (0.7%)	952 (0.2%)
Hyperlipidemia/Hypercholesterolemia	215 (3.3%)	2,025 (0.3%)
Diabetes Mellitus	8 (0.1%)	196 (0.0%)
¹ n (%) unless otherwise specified		

Agreement between ICD-coded and chart-reviewed ACS Ascertainment

Among 619,599 encounters, 5,961 (1.0%) had an ACS ICD code and 4,571 (0.7%) were classified as ACS by chart review. Sensitivity of the ICD codes against the chart-reviewed reference standard was 94.4% (95% confidence interval [CI] 93.7%-95.0%), specificity was 99.73% (CI 99.72%-99.74%), and positive predictive value (PPV) was 72.4% (CI 71.2%-73.5%). (Table 2) Raw agreement was 99.7%, driven mainly by the substantial number of negative cases. Cohen's kappa, which adjusts for these marginal probabilities, was 0.82 (0.81-83).

Table 2
Comparison of ICD Code-Based vs. Chart-Review-Based
Ascertainment of ACS

		ACS diagnosis by chart review		
		Yes	No	Total
ACS diagnosis By ICD code	Yes	(A) 4,314 0.70%	(B) 1,647 0.27%	5,961
	No	(C) 257 0.04%	(D) 613,381 99.00%	613,638
	Total	4,571	615,028	619,599

Following discrepancy review and adjudication (see Reasons for Discrepant Classification), disagreement was confirmed in 1,904 encounters (0.31% of all encounters), comprising 1,647 false positives (0.27%, ICD-coded ACS not confirmed by chart review) and 257 false negatives (0.04%, chart reviewed ACS without an ACS ICD code).

Characteristics by Ascertainment Group

In true-positive cases (n = 4,314), patients had a mean age of 69.9 years and were 60.2% male with 46.9% identifying as White and 14.4% as Hispanic. (Table 3) By comparison, false-positive cases (n = 1,647) were slightly younger (mean 67.2) and less often male (52.6% vs 60.2%). False-negative cases (n = 257) were similar in age to concordant ACS cases (mean 71.0 years) but included a higher proportion of male patients (68.5%) and patients insured by Medicare (65.8%). Concordant non-ACS cases (n = 613,381) were younger (mean 50.9 years), more often female (55.4%), and more often insured by Medicaid (29.2%) or private insurance (34.3%).

Table 3
Demographic Characteristics by ICD Code-Based and Chart Review-Based Ascertainment of ACS

Characteristic	Agree ACS (A) N = 4,314 ²	False positives ICD yes, review no (B) N = 1,647 ²	False negatives ¹ ICD no, review yes (C) N = 257 ²	Agree not ACS (D) N = 613,381 ²
Age				
Mean (SD)	69.9 (15.3)	67.2 (16.4)	71.0 (13.0)	50.9 (20.8)
Median (Q1, Q3)	71.0 (59.0, 82.0)	68.0 (56.0, 80.0)	72.0 (62.0, 80.0)	50.0 (33.0, 67.0)
Min, Max	20.0, 104.0	19.0, 106.0	30.0, 96.0	18.0, 123.0
Sex				
Male	2,597 (60.2%)	867 (52.6%)	176 (68.5%)	273,093 (44.5%)
Female	1,716 (39.8%)	780 (47.4%)	81 (31.5%)	339,885 (55.4%)
Non-Binary	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.0%)
Unknown	1 (0.0%)	0 (0.0%)	0 (0.0%)	401 (0.1%)
Race				
White	2,024 (46.9%)	705 (42.8%)	115 (44.7%)	253,072 (41.3%)
Black or African American	294 (6.8%)	147 (8.9%)	20 (7.8%)	45,364 (7.4%)
Asian	901 (20.9%)	299 (18.2%)	42 (16.3%)	92,688 (15.1%)
Other	885 (20.5%)	415 (25.2%)	67 (26.1%)	195,825 (31.9%)
Unknown/Decline to State	210 (4.9%)	81 (4.9%)	13 (5.1%)	26,432 (4.3%)
Ethnicity				
² n (%) unless otherwise specified.				

Characteristic	Agree ACS (A) N = 4,314 ²	False positives ICD yes, review no (B) N = 1,647 ²	False negatives ¹ ICD no, review yes (C) N = 257 ²	Agree not ACS (D) N = 613,381 ²
Age				
Hispanic/Latino	621 (14.4%)	311 (18.9%)	53 (20.6%)	169,943 (27.7%)
Non-Hispanic/Non-Latino	3,648 (84.6%)	1,317 (80.0%)	203 (79.0%)	434,983 (70.9%)
Unknown/Decline to State	45 (1.0%)	19 (1.2%)	1 (0.4%)	8,455 (1.4%)
Insurance				
Private insurance	998 (23.1%)	330 (20.0%)	34 (13.2%)	210,564 (34.3%)
Medicare	2,557 (59.3%)	957 (58.1%)	169 (65.8%)	188,009 (30.7%)
Medicaid	669 (15.5%)	330 (20.0%)	49 (19.1%)	179,267 (29.2%)
Other	48 (1.1%)	15 (0.9%)	4 (1.6%)	15,813 (2.6%)
Self-Pay/No Insurance	42 (1.0%)	15 (0.9%)	1 (0.4%)	19,728 (3.2%)
Language				
English	3,434 (79.6%)	1,284 (78.0%)	182 (70.8%)	499,090 (81.4%)
Spanish	360 (8.3%)	170 (10.3%)	34 (13.2%)	75,729 (12.3%)
Other	514 (11.9%)	192 (11.7%)	41 (16.0%)	36,944 (6.0%)
Unknown	6 (0.1%)	1 (0.1%)	0 (0.0%)	1,618 (0.3%)
Hypertension	32 (0.7%)	14 (0.9%)	0 (0.0%)	954 (0.2%)
² n (%) unless otherwise specified.				

Characteristic	Agree ACS (A) N = 4,314 ²	False positives ICD yes, review no (B) N = 1,647 ²	False negatives ¹ ICD no, review yes (C) N = 257 ²	Agree not ACS (D) N = 613,381 ²
Age				
Hyperlipidemia/Hypercholesterolemia	148 (3.4%)	41 (2.5%)	10 (3.9%)	2,041 (0.3%)
Diabetes Mellitus	3 (0.1%)	3 (0.2%)	0 (0.0%)	198 (0.0%)
¹ CAD+troponin cases only				
² n (%) unless otherwise specified.				

Reasons for discrepant classification

Of the 1,983 encounters initially classified as discrepant, discrepancy review confirmed 1,904 (96.0%) and reclassified 79 (4.0%) as concordant (Fig. 2). Among the 1,904 confirmed discrepant cases, 1,647 (86.5%) were false positives and 257 (13.5%) were false negatives. Because reviewers could select more than one reason per case, percentages within discrepant categories are not mutually exclusive and do not sum to 100%. Among false positives, the most common reason was a diagnosis of ACS that was suspected but subsequently ruled out: 1,322 cases (80.3%) had been considered for ACS at some point during the encounter but were not confirmed in the ED, including 1,196 (72.6%) in which an initial ED diagnosis of ACS was later revised or ruled out and 250 (15.2%) who presented with a clinical mimic of ACS such as pericarditis, myocarditis, or stress cardiomyopathy. Diagnostic ambiguity accounted for 643 false positives (39.0%), reflecting either documented clinical-team uncertainty about the diagnosis (361, 21.9%) or ambiguous or incomplete documentation (335, 20.3%). A history of CAD that was not active at the time of the encounter accounted for 475 cases (28.8%), including prior ACS coded as a new event (287, 17.4%) and chronic CAD or stable angina miscoded as ACS (247, 15.0%).

Among false negatives, the predominant reason was a coder missing documentation of ACS (221, 86.0%), followed by unclear documentation of ACS by the treating team (100, 38.9%) (Table 4, Fig. 3).

Table 4
Reasons for Discrepancy

	n (%) ¹
False positives (n = 1,647; 86.5% of all discrepant cases)	
A. Diagnostic ambiguity	643 (39.0%)
A1. Primary Team was unclear about ACS / ACS coded despite documented clinical team uncertainty (e.g., ACS may have been considered, but the clinical team did not conclude a final diagnosis)	361 (21.9%)
A2. Unclear documentation by team / Ambiguous or incomplete documentation preventing correct code selection	335 (20.3%)
B. ACS suspected, ruled out later	1,322 (80.3%)
B1. ACS Mimicker / Final cardiac diagnosis a common mimicker of acute coronary syndrome (e.g., pericarditis, myopericarditis, myocarditis, takotsubo (stress) cardiomyopathy, hyperkalemia)	250 (15.2%)
B2. ED Diagnosed ACS - Ruled out later / ED diagnosis of ACS revised or ruled out on inpatient evaluation	1,196 (72.6%)
C. History, not active	475 (28.8%)
C1. History of CAD or angina but not active / Chronic coronary disease or stable angina miscoded as ACS	247 (15%)
C2. History of ACS but not active / Prior history of ACS incorrectly coded as active or new episode of ACS	287 (17.4%)
False negatives (n = 257; 13.5% of all discrepant cases)	
Primary team's documentation for ACS was unclear	100 (38.9%)
Coder missed documentation of ACS	221 (86%)
¹ For each discrepant case, we combined the reasons selected by all reviewers. If the reviewers selected the same reason, it was counted only once. The numbers therefore show how many cases included each reason selected by either reviewer. Percentages were calculated using either the total of false positives or the total of false negatives as the denominator.	

DISCUSSION

In this retrospective cohort of adult ED encounters at a tertiary academic medical center, final hospital ICD codes for ACS showed high agreement with chart-reviewed ACS diagnosis. Sensitivity was 94.4%, specificity was 99.73%, and Cohen's kappa was 0.82. Disagreement was uncommon overall (0.31% of all encounters), and importantly, false negatives, which are the greatest concern in screening

undifferentiated patients presenting for emergency care for ACS, were substantially less frequent than false positives (0.04% vs 0.27%). At this site's volume of approximately 62,000 adult ED encounters annually, this false-negative rate corresponds to roughly 26 cases per year that ICD-based labeling would fail to capture. These findings support the use of ICD codes as training labels for AI models developed to screen ED patients for ACS, including not only the most lethal subtype, STEMI, but also non-ST-segment elevation myocardial infarction (NSTEMI) and unstable angina (UA), which warrant an urgent ECG pathway and subsequent evaluation for potential intervention.

The asymmetry between sensitivity and PPV (94.4% vs 72.4%, respectively) is worth interpreting in the context of AI model development. A model trained on labels with high sensitivity will learn to recognize the clinical signatures of true ACS cases, which is the primary requirement for screening. The lower PPV introduces some positive label noise, which can bias a model toward flagging patterns associated with ACS-adjacent conditions alongside true ACS. This tolerance is specific to the screening test in question: an ECG is inexpensive, rapid, and non-invasive, so the cost of over-flagging is low. A comparable false-positive rate of over-flagging would be less acceptable for a screening pathway triggering invasive or high-cost downstream testing. Lastly, a modest rate of over-flagging is acceptable in this screening application where downstream physician assessment serves as the definitive diagnostic step.

The discrepancies observed also highlight a conceptual consideration specific to training versus deployment. ICD codes are finalized after discharge, which means they cannot serve as real-time inputs to a model operating during an ED encounter. In this study, ICD codes functioned as training-time ground truth only, with a trained model making predictions from information available at arrival early in the visit, such as chief complaint, demographics, vital signs, and initial laboratory values. The validity demonstrated here supports the use of ICD codes to define the training outcome.

The patterns of disagreement we observed help clarify how ICD-based labels differ from treating team diagnoses. The large majority of false positives arose from encounters in which ACS was a part of the diagnostic consideration, either initially suspected and later ruled out or diagnosed as a clinical mimic, rather than from arbitrary coding error. From a modeling perspective, these are informative false positives as they represent patients who clinically warrant evaluation for ACS, which is precisely the population a screening model should flag. A model trained on these labels may trend toward higher sensitivity at the cost of specificity, an error profile that is favorable for ECG screening, where the goal is to avoid missing ACS. False negatives, though far less frequent, were driven by coding gaps rather than by clinical disagreement, suggesting that the small number of missed cases reflects information loss at the coding stage rather than a limitation of the underlying clinical record. Together, these findings suggest that label error in ACS ICD codes is systematic and interpretable, rather than random, and that the dominant direction of error reinforces, rather than undermines, the suitability of these labels for screening-oriented model development.

Our findings are broadly consistent with prior validation work on ACS ICD codes while extending that literature in important ways. A 2014 systematic review of myocardial infarction code validation studies

reported pooled PPVs ranging widely across settings, with most validation conducted in inpatient registries and administrative claims databases rather than in ED cohorts¹⁸. A more recent validation of ICD-10 codes in a Taiwanese insurance claims database reported similar overall accuracy for STEMI and NSTEMI¹⁹. To our knowledge, no comparable contemporary validation has been conducted in a US ED population. Furthermore, our study differs in focusing on a contemporary ED population at a US academic medical center, where the clinical presentation of potential ACS is heterogeneous and coding must reconcile documentation across emergency and inpatient teams. The consistency of our findings with prior work in different settings suggests that ICD codes for ACS perform reasonably well across healthcare contexts, though absolute accuracy likely varies with institutional documentation culture, coder training, and local case mix.

Several aspects of our chart-review approach warrant discussion. Our reference standard was the treating team's diagnosis rather than independent adjudication of true ACS status. Our estimates, therefore, reflect coding fidelity assessing how well ICD codes capture documented diagnoses rather than diagnostic accuracy itself. This is advantageous for our pragmatic purpose, since clinical AI models are trained on documented diagnoses rather than adjudicated truth, making coding-level validity the most relevant target. Validation against objective diagnostic criteria remains an important direction for future work.

This distinction matters because the treating team's documented conclusions reflect contemporary clinical practice, including evolving diagnostic terminology around type 2 myocardial infarction and demand-related ischemia. These events represent a supply-demand mismatch rather than acute atherothrombosis and are conceptually distinct from type 1 ACS under the Fourth Universal Definition of Myocardial Infarction²⁰, recently reclassified as primary myocardial infarction and secondary myocardial infarction in the Fifth Universal Definition²¹. Nonetheless, patients with secondary (formerly type 2) myocardial infarction experience substantial long-term mortality and cardiovascular risk, supporting their identification through the same early screening pathway²². In practice, treating teams frequently document them under ACS-adjacent terminology ("*Type 2 NSTEMI*", "*NSTEMI, suspect demand*", "*ACS, suspect demand*"), which ICD coders typically capture as NSTEMI and thus ACS. We instructed chart reviewers to classify these cases as ACS-present, aligning our reference standard with both the treating team's documentation and contemporary ICD coding practice. Including these cases dilutes the pathophysiologic specificity of ACS by grouping supply-demand events with acute plaque rupture events, but this is consistent with our screening use case. From an ED presentation standpoint, patients who are ultimately diagnosed with secondary (formerly type 2) myocardial infarction warrant the same rapid ECG-based screening pathway as those with primary myocardial infarction (formerly type 1), for which the pathophysiologic distinction is established later with additional clinical context. This classification reflects, rather than departs from, current clinical documentation and coding practice and is aligned with the practical clinical thresholds for early ECG rather than the strict pathophysiologic definition of ACS.

An additional consideration is our use of enrichment criteria for selecting non-ACS encounters for chart review (see Methods, Selection of Encounters for Chart Review). All ACS-ICD-coded encounters were reviewed, but among encounters without an ACS code, we also reviewed those with a CAD ICD code and abnormal troponin. This strategy was designed to capture clinically plausible false negatives, patients for whom ACS could reasonably have been missed, without requiring manual review of the entire non-ACS population.

Several limitations warrant consideration. First, this study was conducted at a single tertiary academic medical center with an active cardiology service and cardiac catheterization laboratory. Variation in coding practices across clinical settings may limit the generalizability of our results to other institutions¹⁸. Second, our ten-year study period (2015–2024) crossed the ICD-9-CM to ICD-10-CM transition in the United States, so our estimates reflect a mix of coding eras rather than any single era's practice. The period also spanned the institutional adoption of high-sensitivity troponin assays^{23–25}, which may have altered which encounters met our enrichment criteria and may have shifted the composition of cases available for review over time. These findings will require reconfirmation as ICD-11 is implemented, particularly given the alignment between ICD-11 and the Fifth Universal Definition.

Beyond these scope considerations, several methodological choices warrant acknowledgment. Concordant cases were not subjected to the same multi-review scrutiny applied to discrepant cases. Although our tiered escalation process minimized misclassification during the initial review, some apparent agreements might have been reclassified upon additional review, a limitation consistent with validation studies that reserve detailed adjudication for discrepant cases. We also relied on enrichment criteria rather than reviewing every non-ACS encounter. Because these criteria required abnormal troponin, they inherently underrepresent UA, which by definition lacks positive cardiac biomarkers. Because UA is estimated to comprise 14% of ACS presentations²⁶, our true false-negative rate is potentially higher than estimated. This limitation primarily affects our ability to detect ICD-negative UA cases since all UA cases with an ACS ICD code were captured by the full chart-review. An ACS encounter with neither an ACS nor CAD ICD code and without abnormal troponin is clinically uncommon.

ICD codes for ACS showed substantial agreement with the treating team's documented diagnosis, with high sensitivity, high specificity, and a low false-negative rate. These findings support ICD-coded ACS as training labels for AI models designed to screen ED patients for an early ECG. Critically, most of the discrepant cases were false positives, many representing patients who warranted ACS consideration, precisely the population a screening tool should flag. In contrast, false negatives were rare and primarily reflected coding gaps. These interpretable discrepancy patterns provide a roadmap for both model design and targeted improvements to cardiovascular coding practices.

METHODS

Study Design and Setting

We conducted a 10-year retrospective cohort study of adult patients presenting to the tertiary academic ED between January 1, 2015 and December 31, 2024. The study compared final hospital ICD codes for ACS with the treating team's final diagnosis of ACS based on a structured chart review of the ED visit, progress notes, consult notes, and discharge summary.

Study Population

We included adult ED visits occurring during the study period, defined as patients ≥ 18 years of age. We excluded encounters with trauma or stroke activations because these patients follow alternative care pathways that take precedence over standard ACS screening workflows.

Ethical approval

and informed consent

This study was reviewed and approved by the Institutional Review Board (IRB) at Stanford University under protocol [IRB#56066]. The study was determined to pose minimal risk to subjects, and a waiver of informed consent was granted because the research involved only a retrospective review of patient data within the health record.

Outcome Definitions

ACS visits were identified using a previously validated set of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes for the subdiagnoses within ACS: STEMI, NSTEMI, and UA^{16,17}. An encounter was classified as ICD-based ACS if an ACS-related ICD code appeared in any of the first 10 final hospital diagnoses for the ED-to-hospital encounter.

The reference standard for this study was the ACS diagnosis documented by the treating team during the index encounter. Trained reviewers identified the treating team's documented diagnosis from the medical record without independently adjudicating its clinical accuracy. This pragmatic reference standard was chosen because models are trained on the diagnoses recorded and coded in structured data rather than on re-adjudicated ground truth. The validity of an ICD code as a training label depends on how faithfully it reflects the documented clinical diagnosis.

Reviewers assessed the discharge summary, ED notes, inpatient progress notes, consultation notes, and other relevant non-billing fields in the EHR to determine whether ACS was a final diagnosis assigned by the treating team. For study purposes, encounters were classified as ACS-present if the treating team documented STEMI, NSTEMI, UA, or secondary (formerly type 2) myocardial infarction/demand ischemia as the final diagnosis²⁰. This definition is intentionally broader than the ICD-based definition, which captures STEMI, NSTEMI, and UA codes. Because treating teams frequently document secondary events under ACS-adjacent terminology that codes capture as NSTEMI, this alignment reflects contemporary documentation and coding practice.

Selection of Encounters for Chart Review

All encounters with an ACS ICD code were selected for chart review. To identify potentially missed ACS among encounters without an ACS ICD code, we also reviewed encounters meeting both of the following enrichment criteria: (1) CAD ICD code and (2) abnormal troponin findings. This enrichment strategy was designed to capture both ACS-ICD code-positive encounters and clinically plausible false-negative encounters without requiring manual review of the entire non-ACS population. Sensitivity and false-negative estimates therefore rest on the assumption that ACS encounters outside these enrichment criteria — those with neither an ACS nor CAD ICD code and without abnormal troponins — were rare. All other encounters were treated as non-ACS in the primary analysis.

Abnormal troponin was defined according to the assay in use during the encounter. This includes conventional troponin I assay and high-sensitivity troponin I assays with abnormal troponin defined by institutional laboratory thresholds.

Chart Review Process

Chart review was performed by nine clinically experienced reviewers who were blinded to the ICD-based diagnosis classification, following established methodological standards for reducing bias in retrospective chart review²⁷. Primary reviewers were required to have at least 1 year of clinical or research experience in emergency care, inpatient medical care, or EHR clinical documentation. All reviewers underwent 3 hours of structured training on the chart review procedure. Training included 1 hour of specified pre-reading and preparation, 2 hours of didactics and chart review methodology, and a performance review of 5 test cases for procedural accuracy before being approved to initiate chart review. A lead reviewer with more than 5 years of experience and specialty training in internal medicine and cardiology supported the reviewers when questions arose. Cases were randomly assigned to reviewers to ascertain presence or absence of the diagnosis of ACS by the treating team. Cases that were unclear were flagged for review by the lead reviewer. If cases remained unclear after lead reviewer assessment, they were referred for additional adjudication by an experienced physician reviewer with no less than 15 years of experience with emergency cardiovascular care. Encounters with a definitive chart-review diagnosis were included in primary analysis.

All cases with different ICD-based and clinically-based ACS diagnosis underwent a secondary review to characterize reasons for disagreement. Two discrepancy review teams independently reviewed each discrepant case. For each case, reviewers could either confirm the discrepancy and select one or more prespecified reasons, or question whether the case was truly discrepant. Those with disagreement among the discrepancy review underwent additional expert adjudication for final classification. Prespecified discrepancy categories included clinical mimics of ACS, prior ACS miscoded as an active event, ambiguity or incompleteness in documentation, and coding omissions or alternative coding choices. Reviewers could select multiple reasons per case.

Patient Characteristics of Chart-Reviewed Cohort

For each visit, we obtained additional details for the patient, including demographic characteristics, insurance type, primary language, and selected cardiovascular comorbidities (history of hypertension, diabetes, or hyperlipidemia).

Statistical Analysis

The primary analysis compared assigned ACS ICD-codes with treating team's diagnosis using a 2 x 2 contingency table. Primary measures of label validity were sensitivity, specificity, PPV, and the false-positive and false-negative rates, each reported with 95% CI. Cohen's kappa was calculated as a secondary agreement measure.^{28,29} Because the large number of concordant non-ACS encounters inflates raw agreement, kappa provides a more informative summary of agreement beyond chance in this setting. Demographic characteristics were summarized across concordant and discrepant groups, and discrepant cases were further categorized by the most likely reasons for disagreement.

Declarations

Data Availability

The data supporting the findings of this study were derived from the EHR at Stanford Health Care and contain protected health information. They are not publicly available because of institutional privacy restrictions and the terms of the approving IRB protocol. De-identified aggregate data may be made available from the corresponding author on reasonable request, subject to institutional review and a data use agreement.

Code Availability

Analyses were performed using R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria). The analytic code used to generate the results reported in this study is available from the corresponding author on reasonable request.

Acknowledgments

This work was supported by the National Institutes of Health and the National Heart, Lung, and Blood Institute, under award numbers R01 HL171403 and T32EB009035, respectively. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

D.K.A. and M.Y.A.B.Y. conceived and designed the study. D.K.A. drafted the manuscript. K.M., A.Ga., H.D., K.R., and A.Gr. designed and performed the statistical analysis. D.K.A., G.B., M.N., V.J.F.M., W.M., J.H., D.L., M.R., and A.C. performed chart review and adjudication. S.D.K. contributed statistical methodology. C.Y. and W.F. provided cardiology domain expertise and critical revision. M.Y.A.B.Y. supervised the study.

and obtained funding. All authors critically revised the manuscript for important edits and intellectual content and approved the final manuscript.

Competing Interests

All authors declare no competing interests.

References

1. Petrella, R.J. The AI Future of Emergency Medicine. *Ann. Emerg. Med.* 84, 139–153 (2024)
2. Taylor, R.A. et al. Impact of artificial intelligence-based triage decision support on emergency department care. *NEJM AI* 2, e2400296 (2025)
3. Almulihi, Q.A. et al. Applications of Artificial Intelligence and Machine Learning in Emergency Medicine Triage: A Systematic Review. *Med. Arch.* 78,198–206 (2024)
4. Song,H., Kim, M., Park, D., Shin, Y. & Lee, J.G. Learning from noisy labels with deep neural networks: a survey. *IEEE Trans. Neural Netw. Learn. Syst.* 34, 8135–8153 (2023).
5. Henderson, T., Shephard, J., & Sundararajan, V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Med Care* 44, 1011–1019 (2006).
6. O'Malley, K.J., et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res.* 40, 1620–1639 (2005)
7. Ramadan, B. et al. Diagnostic Codes in AI Prediction Models and Label Leakage of Same-Admission Clinical Outcomes. *JAMA Netw. Open* 8, e2550454 (2025).
8. Wei, M.Y. et al. Comprehensive review of ICD-9 code accuracies to measure multimorbidity in administrative data. *BMC Health Serv. Res.* 20, 489 (2020).
9. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453 (2019).
10. Banda. J.M., Seneviratne, M., Hernandez-Boussard, T. & Shah, N.H. Advances in electronic phenotyping: from rule- based definitions to machine learning models. *Annu. Rev Biomed. Data Sci.* 1, 53–68 (2018).
11. Agarwal, V. et al. Learning Statistical Models of Phenotypes using a Noisy Labeled Training Data. *J. Am. Med. Inform. Assoc.* 23, 1166–1173 (2016).
12. Rao, S.V. et al. 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes. *Circulation.* 151, e771-e862 (2025)
13. Coyne, C.J., et al. Improving door-to-balloon time by decreasing door-to-ECG time for walk-in STEMI patients. *West J. Emerg. Med.* 16,184–189 (2015).
14. Yiadom, M.Y.A.B, et al. Fallacy of Median Door-to-ECG Time: Hidden Opportunities for STEMI Screening Improvement. *J. Am. Heart. Assoc.* 11, e024067 (2022).
15. Jollis, J.G., et al. Treatment Time and In-Hospital Mortality Among Patients With ST-Segment Elevation Myocardial Infarction, 2018–2021. *JAMA.* 328. 2033–2040 (2022).

16. Bunney. G., et al. Beyond chest pain: Incremental value of other variables to identify patients for an early ECG. *Am. J. Emerg. Med.* 67:70–78 (2023).
17. Yiadom MY, Baugh CW, McWade CM, et al. Performance of Emergency Department Screening Criteria for an Early ECG to Identify ST-Segment Elevation Myocardial Infarction. *J. Am. Heart. Assoc.* 6, e003528 (2017).
18. McCormick, N., Lacaille, D., Bhole, V., Avina-Zubieta, J.A. Validity of myocardial infarction diagnoses in administrative databases: a systematic review. *PLoS One.* 9, e92286. (2014)
19. Tsai, T.Y., et al. Validation of ICD-10-CM Diagnostic Codes for Identifying Patients with ST-Elevation and Non-ST-Elevation Myocardial Infarction in a National Health Insurance Claims Database. *Clin. Epidemiol.* 15,1027–1039 (2023).
20. Thygesen, K., Alpert, J.S., Jaffe, A.S. et al. Fourth Universal Definition of Myocardial Infarction. *J. Am. Coll. Cardiol.* 72, 2231–2264 (2018).
21. Mills, N.L., Newby, L.K., Zaman, S. et al. Fifth Universal Definition of Myocardial Infarction. *J. Am. Coll. Cardiol.* 26, (2026).
22. Chapman, A.R. et al. Long-term outcomes in Patients with Type 2 Myocardial Infarction and Myocardial Injury. *Circulation* 137, 1236–1245 (2018).
23. Chapman, A.R. et al. High-sensitivity cardiac troponin and the Universal Definition of Myocardial Infarction. *Circulation* 141, 161–171 (2020).
24. Ola, O. et al. Clinical Impact of High-Sensitivity Cardiac Troponin T Implementation in the Community. *J Am Coll Cardiol.* 77, 3160–3170 (2021).
25. D’Souza, M. et al. Diagnosis of unstable angina pectoris has declined markedly with the advent of more sensitive troponin assay. *Am. J. Med.* 128, 852–860 (2015).
26. Saxena, M. et al. Variation in ACS patient hospital resource utilization: is it time for advanced NSTEMI risk stratification in the ED? *Am. J. Emerg. Med.* 70, 171–174 (2023).
27. Kaji, A.H., Schriger, D. & Green, S. Looking Through the Retrospectroscope: Reducing Bias in Emergency Medicine Chart Review Studies. *Ann. Emerg. Med* 64, 292–298 (2014).
28. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46 (1960).
29. Landis, J.R. & Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics.* 33, 159–174 (1977).

Figures

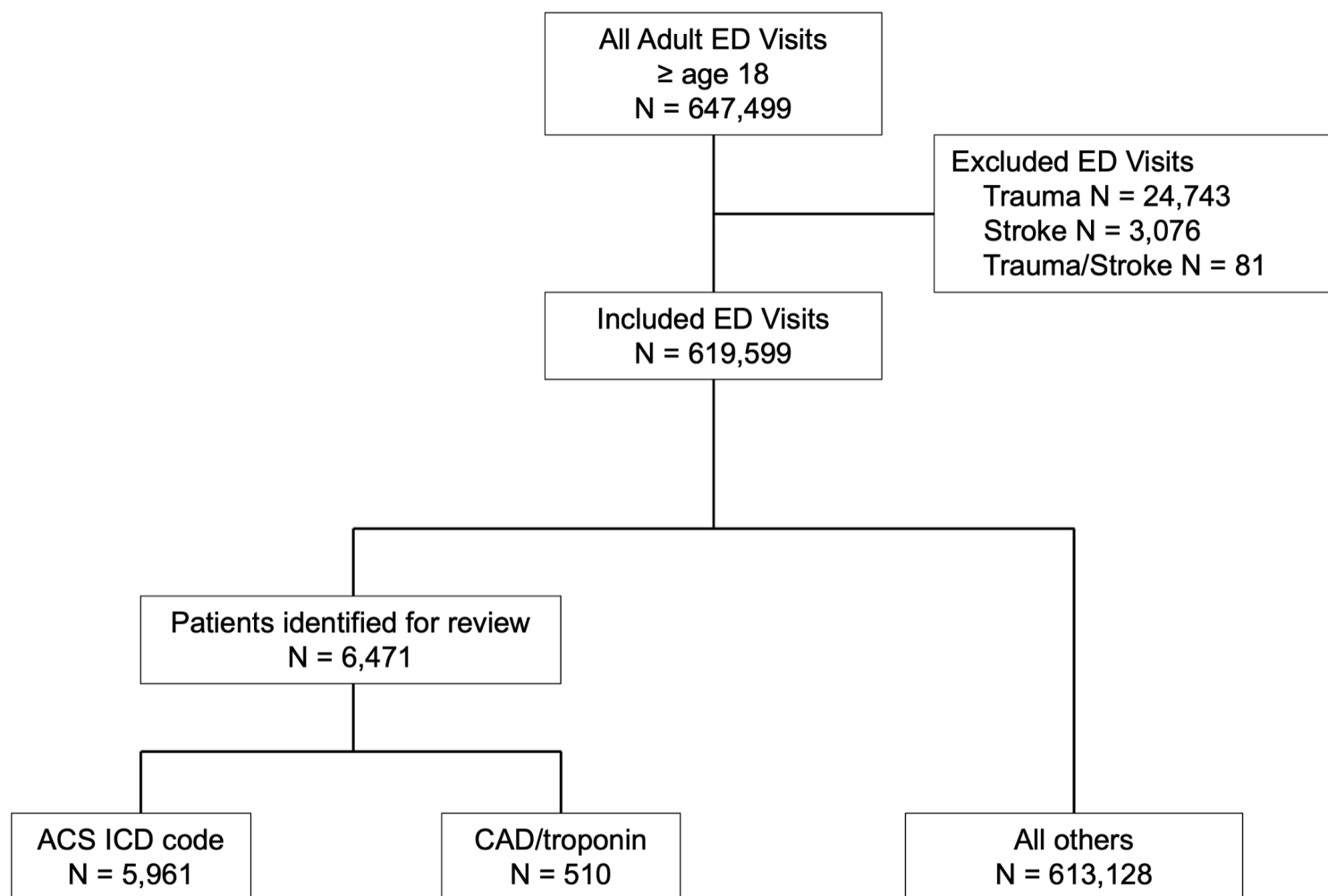


Figure 1

CONSORT diagram

Adult (≥ 18 years) emergency department encounters at a tertiary academic medical center between January 1, 2015 and December 31, 2024. Trauma and stroke activations were excluded because these patients follow alternative care pathways that take precedence over standard acute coronary syndrome (ACS) screening. Encounters were selected for structured chart review if they had an ACS ICD code, or a coronary artery disease (CAD) ICD code with an abnormal troponin result in the absence of an ACS code; the latter criterion enriched for clinically plausible false negatives. All remaining encounters were classified as non-ACS in the primary analysis.

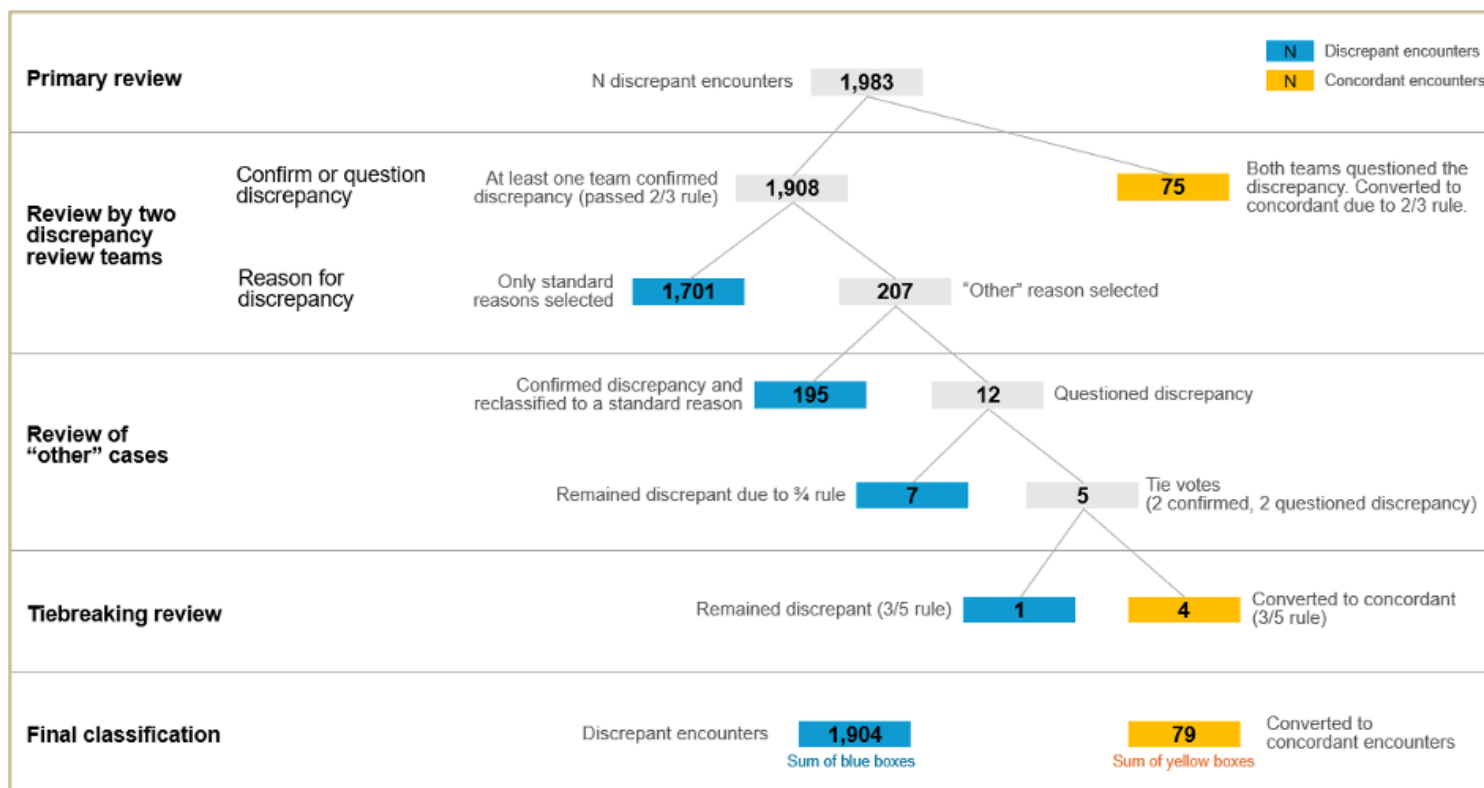


Figure 2

Review of encounters classified as discrepant by primary review.

Each discrepant encounter was independently reviewed by two discrepancy review teams, which could confirm the discrepancy and select one or more prespecified reasons, or question whether the case was truly discrepant. Final classification was determined by majority agreement across the primary review and the two discrepancy reviews, with tied cases referred for expert adjudication. Of 1,983 encounters initially classified as discrepant, 1,904 (96.0%) were confirmed and 79 (4.0%) were reclassified as concordant.

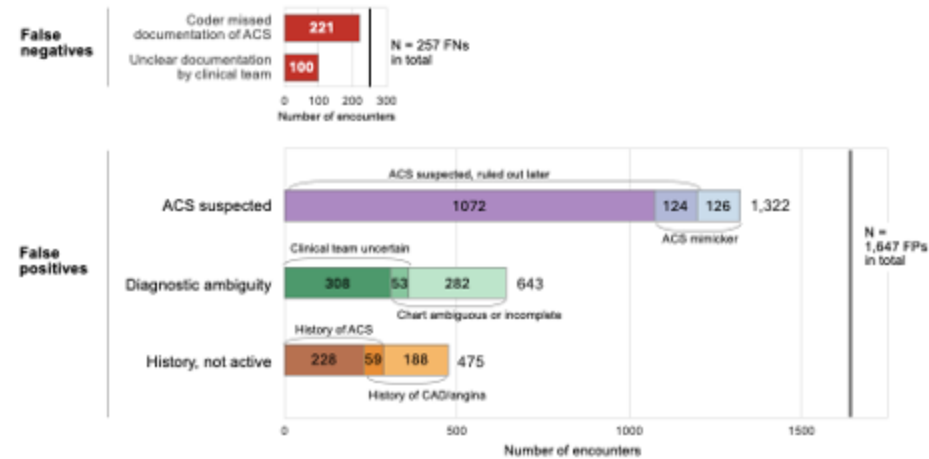


Figure 3

Reasons for false positive and false negative encounters.

Reasons for discrepant classification among confirmed false-positive (n=1,647) and false-negative (n=257) encounters. Reviewers could select more than one reason per encounter; percentages are therefore not mutually exclusive and do not sum to 100%. Percentages are calculated using the total number of false positives or false negatives as the denominator.