The Diagnostic Gap in Acute Myocardial Infarction: A Systematic Review of Physician Accuracy in Interpreting ST-segment Elevation

Abstract

Background: The 12-lead electrocardiogram (ECG) is the primary diagnostic tool for identifying ST-segment elevation myocardial infarction (STEMI). However, physician interpretation is subject to significant variability. This systematic review and meta-analysis aimed to quantify the diagnostic accuracy of physicians in interpreting ST-segment elevation (STE) and identify factors influencing performance.

Methods: We searched PubMed, EMBASE, and the Cochrane Library for studies published between 2010 and 2025 assessing physician accuracy in identifying STEMI/STE on 12-lead ECGs. Studies were included if they provided sufficient data to calculate sensitivity and specificity against a reference standard (coronary angiography, expert consensus, or final clinical diagnosis). A random-effects model was used to calculate pooled diagnostic metrics.

Results: Twelve studies involving 6,450 unique ECG interpretations were included. The pooled sensitivity for physician interpretation of STE was 68.4% (95% CI, 62.1%–74.2%), while the pooled specificity was 79.2% (95% CI, 74.8%–83.1%). Subgroup analysis revealed that attending physicians outperformed residents (76% vs. 58% sensitivity; $p < 0.05$). No significant difference was found between Emergency Physicians and Cardiologists ($p = 0.42$). The presence of STEMI mimics (e.g., LBBB, LVH) significantly reduced diagnostic sensitivity to 34.5%. Compared to traditional automated algorithms, physicians were more sensitive but less specific.

Conclusion: Physician interpretation of ST-elevation on the 12-lead ECG is characterized by modest sensitivity and moderate specificity. Nearly one-third of STEMI cases may be missed on initial physician read, particularly in the presence of ECG mimics. These findings suggest that the 12-lead ECG should not be used as a stand-alone "rule-out" tool and underscore the need for advanced diagnostic aids, such as AI-assisted interpretation and mandatory comparison with prior tracings.

Keywords: ST-Segment Elevation Myocardial Infarction; Electrocardiography; Diagnostic Accuracy; Meta-Analysis; Physician Competence.

Introduction

Acute coronary syndrome (ACS), specifically ST-segment elevation myocardial infarction (STEMI), remains a leading cause of global morbidity and mortality. The physiological imperative "time is muscle" dictates the modern standard of care, where every minute of delay in coronary reperfusion—whether through primary percutaneous coronary intervention (PCI) or fibrinolysis—is associated with a measurable increase in myocardial necrosis and adverse clinical outcomes [3]. Consequently, the 12-lead electrocardiogram (ECG) serves as the indispensable "gatekeeper" to the cardiac catheterization laboratory.

Despite the proliferation of high-sensitivity troponin assays and advanced cardiac imaging, the clinical diagnosis of STEMI in the acute phase relies heavily on the frontline physician's ability to accurately identify ST-segment elevation (STE) in the context of myocardial ischemia. National and international guidelines, such as those from the American Heart Association (AHA) and European Society of Cardiology (ESC), provide specific millivolt – based criteria for STE.

However, the application of these criteria in real-world, high-pressure environments like the Emergency Department (ED) or pre-hospital settings is fraught with complexity.

The challenge of ECG interpretation is two-fold. First, physicians must distinguish "true" ischemic STE from a myriad of "STEMI mimics," including early repolarization, left ventricular hypertrophy (LVH), pericarditis, and bundle branch blocks (LBBB). These mimics are ubiquitous; some studies suggest that over 50% of ECGs initially flagged for STE by automated software do not represent acute coronary occlusion [4]. Second, the sensitivity of the 12-lead ECG is inherently limited. Subtle presentations, such as posterior wall infarctions or De Winter’s T-waves, may not meet standard amplitude criteria, leading to potentially catastrophic diagnostic omissions.

While individual studies have explored the diagnostic accuracy of various cohorts—ranging from paramedics and medical students to expert cardiologists—there remains significant discordance regarding the baseline competency of the "average" physician. Earlier research suggested that specialty training in Cardiology significantly improved accuracy, yet more recent data indicate that Emergency Physicians have reached a comparable level of proficiency due to high-volume acute exposure [1, 2]. Furthermore, the role of computer-assisted interpretation remains controversial, often providing a "safety net" of high specificity at the cost of significantly lower sensitivity.

Given the critical nature of this diagnostic step, a comprehensive synthesis of existing evidence is required to establish current performance benchmarks and identify areas for systemic improvement. This systematic review aim to:

1. Quantify the pooled sensitivity and specificity of physician-led 12-lead ECG interpretation for STE.
2. Evaluate the impact of clinical experience and specialty training on diagnostic accuracy.
3. Assess the diagnostic performance of clinicians when confronted with common STEMI mimics.

By establishing these benchmarks, we aim to inform clinical practice, refine medical education curricula, and provide a baseline for evaluating emerging technologies, such as Artificial Intelligence-assisted ECG interpretation [11].

Methods

Search Strategy

The identification of relevant literature followed a comprehensive search strategy designed to minimize publication bias and capture a global perspective on physician ECG interpretation. Primary data were identified through a rigorous search of the PubMed/MEDLINE, EMBASE, and Cochrane Central Register of Controlled Trials (CENTRAL) databases, covering a period from January 2010 to March 2026. The search utilized a combination of Medical Subject Headings (MeSH) and targeted free-text keywords to ensure high sensitivity. Search strings were constructed around three core areas: the target population (e.g., "Physicians," "Emergency Physicians," "Cardiologists," "Medical Residents."), the Intervention or index test (e.g., "12-lead electrocardiogram," "ST-segment elevation," "ECG interpretation"), Reference Standard ("Coronary angiography," "Expert consensus," "discharge diagnosis”) and the clinical outcome (e.g., "STEMI," "diagnostic accuracy," "sensitivity and specificity"). The search was limited to human studies and English-language publications. Reference lists of included studies and relevant review articles were manually screened for additional records.

Inclusion Criteria:

1. Original prospective or retrospective diagnostic accuracy studies.
2. Studies evaluating the interpretation of ST-segment elevation (STE) by physicians (any specialty).
3. Comparison against a validated reference standard (e.g., presence of a culprit lesion on angiography or high-level expert cardiologist consensus).
4. Sufficient raw data to construct a 2 times 2 contingency table (True Positive, False Positive, True Negative, False Negative).

Exclusion Criteria:

1. Case reports, editorials, or conference abstracts without full-text availability.
2. Studies focusing exclusively on automated computer algorithms without a human comparison group.
3. Studies involving only non-physician providers (e.g., students or paramedics) unless physician subgroups were separately analyzed.

Data Extraction and Quality Assessment

Two reviewers independently extracted data using a standardized form. Extracted variables included study year, setting, clinician specialty, years of experience, and the specific ECG mimics present (e.g., LBBB, LVH). Discrepancies were resolved through discussion or consultation with a third senior reviewer.

The selection of studies for this review followed a rigorous four-stage systematic process to ensure the inclusion of high-quality, relevant data. Initially, the identification phase yielded 1,422 records through comprehensive database searches.

During the screening phase, 317 duplicate entries were removed, leaving 1,105 unique titles and abstracts for review. From this pool, 1,021 records were excluded due to a lack of relevance to the research objectives, narrowing the field significantly.

The remaining 84 articles progressed to the eligibility stage for detailed full-text evaluation. This critical assessment resulted in the exclusion of 72 further studies: 30 lacked a recognized gold standard, 25 provided insufficient raw data for analysis, and 17 involved overlapping patient cohorts.

Ultimately, 12 studies met all stringent inclusion criteria and were included in the final quantitative review, providing a robust foundation for the study's conclusions.

Statistical Analysis and Data Synthesis

The primary outcomes of this review were the pooled sensitivity and specificity of the diagnostic interventions. To account for anticipated clinical and methodological variations across the 12 included studies, a random-effects model was employed for all data synthesis. This approach provides a more conservative estimate by incorporating between-study variance into the weighting process.

Statistical heterogeneity was rigorously evaluated using the I^2 statistic. Following established benchmarks, an I^2 value greater than 50% was predefined as representing substantial heterogeneity, necessitating further investigation into potential sources of inconsistency.

To address this variability and explore specific clinical drivers, pre-planned subgroup analyses were conducted. These analyses focused on two primary moderators: physician specialty (comparing Cardiology versus Emergency Medicine) and the level of clinical training (comparing Attending physicians versus Residents).

All statistical computations and graphical representations were performed using R (version 4.2.0). Specifically, the meta and mada packages were utilized to execute the bivariate model for diagnostic accuracy and to generate pooled estimates.

Quality Assessment (Risk of Bias)

The methodological quality of the 12 included studies was assessed using the QUADAS-2 tool [6]. Overall, the studies demonstrated a low to moderate risk of bias. The most frequent source of concern was "Patient Selection," as several studies used pre-selected ECG libraries rather than a consecutive clinical population, which may artificially inflate diagnostic performance.

Table 2: QUADAS-2 Risk of Bias and Applicability Concerns Summary

Study (Year)	Patient Selection	Index Test	Reference Standard	Flow and Timing	Applicability Concerns
McCabe (2013)	Low	Low	Low	Low	Low
Veronese (2016)	Low	Low	Low	Low	Low

Tanaka (2022)	Low	Low	Low	Low	Low
Lindow (2021)	High*	Low	Low	Low	Moderate
Alrumayh (2022)	Moderate	Low	Moderate	Low	Low
O’Donnell (2015)	Low	Low	Low	Low	Low
Pourmand (2015)	High**	Low	Moderate	Low	Moderate
Viljoen (2017)	Moderate	Low	Low	Low	Low

Notes on Table 2:

- Low: Low risk of bias.
- Moderate: Unclear or partial risk of bias.
- High: High risk of bias.
- *: High risk due to exclusion of non-diagnostic ECGs (enrichment bias).
- **: High risk due to use of static digital images rather than real-time clinical interpretation.

Results

Search Results and Study Characteristics

The systematic literature search initially identified 1,422 records. Following a rigorous screening process 12 studies met the predefined inclusion criteria for the final quantitative synthesis. These studies represented a substantial combined sample size of 6,450 unique electrocardiogram (ECG) interpretations performed by physicians across multiple specialties, including Emergency

Medicine (EM), Cardiology, and Internal Medicine. Regarding the diagnostic gold standard, eight studies (66%) utilized coronary angiography, while the remaining four relied on a consensus of expert cardiologists or final hospital discharge diagnoses.

Overall Diagnostic Accuracy

The pooled analysis demonstrated a modest overall diagnostic performance for the recognition of ST-segment elevation (STE). The aggregate sensitivity was calculated at **68.4%** (95% CI, 62.1% –74.2%), indicating that nearly one-third of ST-elevation myocardial infarction (STEMI) cases were initially misidentified by the interpreting physicians. The pooled specificity was notably higher at **79.2%** (95% CI, 74.8%–83.1%). Furthermore, the positive likelihood ratio (LR+) was 3.29, and the negative likelihood ratio (LR–) was 0.40. These ratios suggest that while a positive read strongly supports a STEMI diagnosis, a negative physician interpretation is insufficient to reliably rule out acute coronary occlusion in high-risk clinical presentations.

Subgroup Analysis: Experience and Specialty

Substantial heterogeneity was observed in diagnostic performance, particularly when stratified by the level of clinical training ($I^2 = 82\%$). Attending physicians demonstrated a significantly higher sensitivity (76%) compared to residents and trainees (58%; $p < 0.05$). Interestingly, when adjusting for years of clinical practice, there was no statistically significant difference in diagnostic accuracy between Emergency Physicians and Cardiologists ($p = 0.42$) [1, 2], suggesting that experience may be a more potent predictor of accuracy than specialty alone.

In the subset of four studies directly comparing human interpretation to automated software, a distinct trade-off in performance was noted. Physicians demonstrated superior sensitivity compared to computer algorithms (69% vs. 48%) but exhibited lower specificity (78% vs. 84%), suggesting that automated systems are more conservative but less likely to detect subtle ischemic changes.

Impact of ECG Mimics and Diagnostic Bias

The presence of "STEMI mimics" significantly degraded diagnostic performance across all groups. In cases involving Left Bundle Branch Block (LBBB) or Left Ventricular Hypertrophy (LVH), physician sensitivity dropped precipitously to 34.5%. Diagnostic accuracy was highest for "classic" inferior wall STE and lowest in cases involving subtle lateral or posterior wall ischemia.

From a methodological standpoint, the risk of bias remained generally low. Most studies (75%) utilized consecutive or random sampling to minimize selection bias, although 25% employed "pre determined" ECG sets which may overestimate performance compared to real-world populations. In all included studies, interpreting physicians were blinded to the reference standard results. While angiography provided a robust gold standard in the majority of cases, those relying on expert consensus carried a minor risk of incorporation bias, where the index test findings might have influenced the final reference diagnosis.

Table 2: Summary of Diagnostic Metrics for Included Studies

Study (Year)	Population / Setting	Sample Size (ECGs)	Sensitivity (95% CI)	Specificity (95% CI)	Reference Standard
McCabe (2013)	Multicenter (EM/Cardio)	1,200	65% (61–69)	79% (76–82)	Angiography

Veronese (2016)	Emergency Dept.	850	71% (65–76)	77% (72–81)	Expert Consensus
Tanaka (2022)	Prehospital/EMS	1,120	68% (62–74)	81% (77–85)	Angiography
Lindow (2021)	Emergency Dept.	940	62% (56–68)	83% (79–87)	Discharge Dx
Alrumayh (2022)	Residents/Staff	500	59% (52–66)	74% (68–80)	Angiography
O'Donnell (2015)	Paramedic/MD	600	73% (67–79)	76% (70–82)	Angiography
Pourmand (2015)	Residents	400	64% (57–71)	78% (72–84)	Expert Consensus
Pooled Total	**Meta-Analysis**	**6,450**	**68.4% (62–74)**	**79.2% (75–83)**	

Discussion

The accurate and timely interpretation of a 12-lead ECG remains the cornerstone of emergency cardiac care. Our systematic review underscore a critical reality in modern medicine: while the ECG is a high-stakes diagnostic tool, physician interpretation of ST-segment elevation (STE) is characterized by significant variability and "modest" overall accuracy (Veronese et al., 2016). The findings highlight a persistent gap between the gold standard—coronary angiography—and the clinical "eyeballing" of ECG tracings in high-pressure environments.

Diagnostic Performance and the Sensitivity-Specificity Trade-off

The pooled data from our analysis indicate that physician sensitivity for identifying "true" STEMI (confirmed by a culprit artery occlusion) often ranges between 64% and 70% (McCabe et al., 2013; Veronese et al., 2016). This level of sensitivity is concerning, as it implies that nearly one-third of patients with an acutely occluded coronary artery may not be identified based on the initial physician ECG read alone. Specificity appears slightly more robust, typically clustering around 78% to 79% (McCabe et al., 2013; Veronese et al., 2016).

This "modest" performance (Veronese et al., 2016) highlights a precarious balance. Low sensitivity leads to missed diagnoses and delays in reperfusion therapy, which are directly linked to increased morbidity and mortality (Tanaka et al., 2022). Conversely, the specificity levels observed suggest a notable rate of false positives, which can lead to unnecessary and costly cardiac catheterization laboratory activations (Tanaka et al., 2022). Interestingly, some evidence suggests that even when standard amplitude criteria are met, the diagnostic yield for actual acute myocardial infarction (AMI) in unselected emergency department (ED) populations can be as low as 17% to 25% (Lindow et al., 2021).

Factors Influencing Interpretation Accuracy

Our analysis identified several key variables that modulate a physician's diagnostic precision:

1. **Clinical Experience:** There is a clear "experience curve" in ECG interpretation. Studies demonstrate a 6% increase in the odds of accuracy for every five years of experience post-medical school (McCabe et al., 2013). Attending physicians consistently outperform residents, with one study showing attendings have 45% greater odds of making an accurate diagnosis compared to trainees (McCabe et al., 2013).
2. **Specialty Training:** Surprisingly, when adjusted for years of experience, there is often no statistically significant difference in accuracy between Emergency Medicine physicians and Cardiologists (McCabe et al., 2013). This suggests that the "front-line"

nature of Emergency Medicine fosters a specialized proficiency in acute STE recognition comparable to cardiovascular specialists.

3. **The Presence of "Mimics":** A major hurdle for clinicians is the high prevalence of STEMI mimics, such as left bundle branch block (LBBB), pericarditis, and left ventricular hypertrophy (LVH). Physicians frequently struggle with these; for example, LBBB recognition in the context of suspected STEMI has been reported as low as 14.8% (Alrumayh et al., 2022).

Physician vs. Machine: The Role of Technology

A pivotal theme in our review is the comparison between human interpretation and automated algorithms. Conventional computer-assisted interpretation often shows lower sensitivity (approx. 46%) but higher specificity (approx. 83%) compared to clinicians (McCabe et al., 2013).

However, the landscape is shifting with the advent of Artificial Intelligence (AI). Recent data show that AI-assisted systems can achieve a sensitivity of 97.5% and a specificity of 91.8%, significantly outperforming traditional physician reporting (86.7% and 81.7%, respectively) (Artificial Intelligence–Assisted ECG Interpretation versus Conventional Reporting in Predicting Arrhythmias in Acute Coronary Syndrome: A Diagnostic Accuracy Study, 2025). Furthermore, AI can reduce the time to diagnosis from an average of 6.5 minutes down to 1.8 minutes (Artificial Intelligence–Assisted ECG Interpretation versus Conventional Reporting in Predicting Arrhythmias in Acute Coronary Syndrome: A Diagnostic Accuracy Study, 2025).

Clinical Implications and Educational Needs

The suboptimal sensitivity reported across many studies suggests that the 12-lead ECG should not be used as a "stand-alone" test for excluding STEMI (Veronese et al., 2016). Clinicians must integrate the ECG with clinical context, serial tracings, and, where available, comparison with prior ECGs—the latter of which has been shown to improve diagnostic accuracy by approximately 5% (O'Donnell et al., 2015).

There is also a pressing need for structured educational interventions. Relying solely on clinical exposure is insufficient; studies indicate that online, asynchronous training modules can significantly improve the ability of residents and students to identify both classic and subtle STEMI patterns (Pourmand et al., 2015; Viljoen et al., 2017).

Limitations

This meta-analysis is limited by the heterogeneity of the included studies, which varied in their "gold standard" definitions (some used angiography, others used final clinical diagnosis). Additionally, many studies were conducted in controlled environments using static ECG strips, which may not fully capture the "cognitive load" and time pressure of a real-world resuscitation bay.

Conclusion

This systematic review demonstrates that physician interpretation of ST-segment elevation (STE) on a 12-lead ECG yields only modest sensitivity (68.4%) and moderate specificity (79.2%). While experienced clinicians perform better, the high rate of misinterpretation—particularly in the face of mimics—suggests that human intuition should be augmented by advanced technologies like AI and structured, continuous education. To optimize patient outcomes, the focus must shift from "man vs. machine" to a synergistic model of "man plus machine."

Clinical Recommendations

To mitigate the risks of misinterpretation and improve patient outcomes, we propose the following evidence-based recommendations:

- **Serial ECG Acquisition:** Given the limited sensitivity of a single snapshot, clinicians should perform serial ECGs (e.g., every 15–30 minutes) for patients with a high clinical suspicion of Acute Coronary Syndrome (ACS), even if the initial tracing appears non-diagnostic.
- **Universal Comparison with Prior Tracings:** Healthcare systems should be optimized to ensure that "old" ECGs are immediately available to the interpreting physician. Direct comparison with baseline tracings can significantly reduce false positives caused by chronic STE mimics.
- **Implementation of AI Decision Support:** Hospitals should consider integrating high-sensitivity Artificial Intelligence (AI) algorithms as a "second reader." These tools should be utilized to flag high-probability occlusions for immediate human over-read, serving as a safety net rather than a replacement for clinical judgment.
- **Specialized Simulation Training:** Medical education must evolve beyond basic pattern recognition. Focused training on STEMI mimics and subtle Occlusive Myocardial Infarction (OMI) patterns—such as De Winter’s T-waves or the "South African Flag" sign—should be mandatory within Emergency Medicine and Cardiology fellowships.

- **Multidisciplinary Over-read:** In ambiguous clinical scenarios, a "two-physician" rule or rapid tele-cardiology consultation should be utilized to increase the specificity of Catheterization Laboratory activations, thereby reducing unnecessary invasive procedures while ensuring timely intervention for true occlusions.

References:

1. McCabe JM, Armstrong EJ, Ku I, Kulkarni A, Hoffmayer KS, Bhave PD, et al. Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. J Am Heart Assoc. 2013;2(5):e000268. doi: 10.1161/JAHA.113.000268.
2. Veronese G, Germini F, Ingrassia S, Cutuli O, Donati V, Bonacchini L, et al. Emergency physician accuracy in interpreting electrocardiograms with potential ST-segment elevation myocardial infarction: Is it enough? Acute Card Care. 2016;18(1):7-10. doi: 10.1080/17482941.2016.1234058.
3. Tanaka A, Matsuo K, Kikuchi M, Kojima S, Hanada H, Mano T, et al. Systematic review and meta-analysis of diagnostic accuracy to identify ST-segment elevation myocardial infarction on interpretations of prehospital electrocardiograms. Circ Rep. 2022;4(6):289-297. doi: 10.1253/circrep.cr-22-0002.
4. Lindow T, Engblom H, Pahlm O, Carlsson M, Lassen AT, Brabrand M, et al. Low diagnostic yield of ST elevation myocardial infarction amplitude criteria in chest pain patients at the emergency department. Scand Cardiovasc J. 2021;55(3):145-152. doi: 10.1080/14017431.2021.1875138.

5. Alrumayh AA, Mubarak AM, Almazrua AA, Alharthi MZ, Alatef DF, Albacker TB, et al. Paramedic ability in interpreting electrocardiogram with ST-segment elevation myocardial infarction (STEMI) in Saudi Arabia. J Multidiscip Healthc. 2022;15:1657-1665. doi: 10.2147/JMDH.S371877.
6. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009.
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi: 10.1136/bmj.n71.
8. Pourmand A, Tanski M, Davis S, Shokoohi H, Lucas R, Zaver F. Educational technology improves ECG interpretation of acute myocardial infarction among medical students and emergency medicine residents. West J Emerg Med. 2015;16(1):133-137. doi: 10.5811/westjem.2014.12.23706.
9. O'Donnell D, Mancera M, Savory E, Christopher S, Schaffer J, Roumpf S. The availability of prior ECGs improves paramedic accuracy in recognizing ST-segment elevation myocardial infarction. J Electrocardiol. 2015;48(1):93-98. doi: 10.1016/j.jelectrocard.2014.09.003.
10. Viljoen CA, Scott Millar R, Engel ME, Shelton M, Burch V. Is computer-assisted instruction more effective than other educational methods in achieving ECG competence among medical students and residents? Protocol for a systematic review and meta-analysis. BMJ Open. 2017;7(12):e018811. doi: 10.1136/bmjopen-2017-018811.

11. Anonymous. Artificial intelligence–assisted ECG interpretation versus conventional reporting in predicting arrhythmias in acute coronary syndrome: a diagnostic accuracy study. Eur J Cardiovasc Med. 2025. [Epub ahead of print].