Reprint from The Review of DIABETIC STUDIES

The Review Of
DIABETIC
STUDIES OPEN ACCESS

Diagnostic Accuracy And Clinical Safety Of Artificial Intelligence For Detecting And Flagging Critical Laboratory Results: A Systematic Review

Abdulmajeed Ahmad Alsofiany[1], Yousef Mohammed Alqarih[2], Ali Jaber Mohammad Alyami[3], Ibrahim Hussein Ibrahim Alfaqih[4], Mohammed Ali Ali Mushahhim[5], Faisal Rashed Alazizi[6], Saud Ayed Alharthi[7], Alaa Zaed Alzaedi[8]

[1]Family Medicine Consultant, Taif Health Cluster, Taif, Saudi Arabia, Email: aalsofiany@moh.gov.sa
[2]Lab Technician, East Nakhb Health Center, Taif Health Cluster, Email: yalqarih@moh.gov.sa
[3]Laboratory Technician, NAJRAN, Badr Al Janoub Hospital, Najran Health Cluster, Email:Aljaalyami@moh.gov.sa
[4]Paramedicine Specialist, Flight and Critical Care Paramedic, Supervision at Emergency, Disaster, Crisis and Medical Transport Department in AlBaha Health Cluster, Email: ihalfaqih@moh.gov.sa
[5]Specialist-Public Health, Community Health, Medical Transport Pathway Lead in Al-Baha Health Cluster, Email: mmushahhim@moh.gov.sa
[6]Health Assistant Health Security, Inside Saudi Arabia, Taif, Email:Falmutairi204@moh.gov.sa
[7]Otorhinolaryngology Head and Neck Surgery Consultant, Department of Otorhinolaryngology Head and Neck Surgery, King Abdul Aziz Specialist Hospital, Taif, Email:Saayalharthi@moh.gov.sa
[8]Physical Medicine and Rehabilitation Registrar, Rehabilitation Center King Abdulaziz Specialist Hospital, Taif, Saudi Arabia, Email: Azalzaedi@moh.gov.sa

Abstract

Background: Delayed or missed critical laboratory results remain a preventable source of patient harm. The practice of flagging them dates to George Lundberg's original 1972 definition of a critical value as a result "life-threatening unless something is done promptly" (Lundberg, 1972), and the manual call-and-document workflow built around that definition has changed remarkably little in the half-century since. Artificial intelligence (AI) and machine learning (ML) tools are now increasingly proposed to detect or flag these results faster and more reliably than existing threshold-based alerting.

Objective: To evaluate the diagnostic accuracy and clinical safety of AI/ML systems for detecting or flagging critical laboratory and physiological results, and to examine how far the regulatory and implementation evidence has kept pace with the technical claims.

Methods: We searched PubMed/MEDLINE, Embase, CINAHL, Scopus, Web of Science, IEEE Xplore, and the Cochrane Library from [start date] to [search date], following PRISMA 2020 and its diagnostic test accuracy extension, PRISMA-DTA. Two reviewers screened records independently. [Risk of bias was assessed with QUADAS-2, supplemented by the QUADAS-AI framework, and certainty with GRADE.] Because study designs and reference standards varied too widely to pool, findings were synthesised narratively by critical-value category.

Results: Of 89 records, 11 studies were included. AI-enabled electrocardiography for detecting severe hyperkalemia was the most mature application, with two independent studies. Hypokalemia detection, anemia detection, and glucose-abnormality prediction were each represented by one to three studies, with wider and less consistent accuracy ranges. No included study reported a clinical safety outcome such as notification time, missed-result rate, or clinician override behaviour, and none tested for automation bias.

Conclusions: AI and machine learning can detect several types of critical result with accuracy approaching or matching laboratory confirmation, most convincingly for ECG-based hyperkalemia detection. Whether these tools improve, or compromise, the safety of the notification pathway itself is currently unknown, because almost no study in this literature has measured it.

Keywords: artificial intelligence; machine learning; critical values; diagnostic accuracy; patient safety; electrocardiography; clinical alerts; regulatory science.

1. Introduction

1.1 The weight of a single number

A critical laboratory result is one of the few moments in medicine where a single number can outrank almost everything else happening around a patient. A potassium of 7.2. A glucose of 25. A hemoglobin that has quietly dropped overnight without anyone noticing the bleed. A blood culture flagged positive at three in the morning. George Lundberg gave this category of result its name in 1972, defining a critical, or "panic," value as one representing a pathophysiologic state so far from normal that it is likely to be life-threatening unless something is done promptly, and for which some corrective action can actually be taken (Lundberg, 1972). That definition has survived, almost word for word, for more than fifty years, which is itself worth pausing on. Medicine has reinvented nearly everything else about how it finds and treats disease in that time. The basic idea of a critical value, and the belief that it deserves a different kind of handling than an ordinary result, has not needed reinventing.

What has changed is how seriously that handling is taken, and how much is now known about what happens when it fails. A landmark survey of 623 institutions participating in the College of American Pathologists' Q-Probes program found that the overwhelming majority of physicians, 95%, considered notification of a critical result genuinely useful to how they managed a patient, and that two-thirds of critical results prompted some real change in treatment (Howanitz et al., 2002). That is not a marginal effect. It means that in roughly two cases out of three, the phone call from the laboratory was not a formality, it was the moment the plan changed.

1.2 Where the current system still breaks

The workflow built to deliver that phone call, though, is still surprisingly manual, and has been for the whole of its existence. A technologist spots the flagged value, checks it against a locally defined threshold list, and calls someone. Accreditation bodies such as the College of American Pathologists and the Joint Commission require that a critical value reach a responsible clinician within a defined window, often cited as 30 minutes for 90% of results, and yet even closed-loop systems built specifically to hit that target continue to report real-world notification times that drift past it, particularly overnight, during shift changes, or when the first person called doesn't pick up. Howanitz and colleagues' own survey found wide variation across laboratories in how rigorously critical values were even tracked, let alone delivered on time (Howanitz et al., 2002). Half a century after Lundberg coined the term, the weak link in the system is still often the same one: a human being, at the end of a phone line, who may or may not answer.

This is not a criticism of the people doing that job. It is an observation about the structure of the task. Verifying a flagged result, locating the right clinician, and documenting the call all take time that a busy laboratory, especially overnight, does not always have to spare. And the threshold lists themselves carry an old, known weakness: they are fixed numeric cutoffs, chosen once by a laboratory's own staff, often with limited external validation, which means they can be too sensitive for some patients and not sensitive enough for others (Howanitz et al., 2002).

1.3 Two ways artificial intelligence has entered the picture

Artificial intelligence has started to enter this workflow from more than one direction, and it is worth being precise about which direction, because the literature tends to blur them together in a way that makes the evidence look more unified than it is.

The first approach keeps the laboratory result as the primary signal but tries to see the critical state coming before the threshold is actually crossed, using the surrounding pattern of a patient's other routine labs to predict an approaching mortality risk or deterioration (Shimada et al., 2023). The second approach abandons the laboratory draw altogether for certain tasks and substitutes a different, faster signal. AI-enabled electrocardiography has been developed to infer dangerous potassium levels non-invasively from the shape of the ECG waveform itself, sometimes minutes before a blood sample would even reach the lab (Lin et al., 2020; An et al., 2024).

What connects both is a shared bet, even though the mechanisms are completely different: that a machine can either be faster than the current callback chain, more sensitive than a fixed numeric cutoff, or both at once.

1.4 What the evidence shows so far

The evidence so far is genuinely encouraging in places, and it would be a disservice to this literature to undersell that. AI-ECG models for detecting severe hyperkalemia have reported strong discrimination in independent studies (Galloway et al., 2019; Lin et al., 2020). Studies that sit outside this review's synthesis point in a similar direction and are cited here as background only: a deep learning model for detecting hyperkalemia and hypokalemia from the ECG has been tested in a separate external validation cohort (An et al., 2024), and machine learning models have been built to predict hypoglycemia in hospitalized patients from routine electronic health record data (Mathioudakis et al., 2021; Yang et al., 2022).

1.5 The safety side nobody has measured

The safety side of the question has received far less attention than the accuracy side, and this is where the real gap in the literature sits, and where this review spends most of its energy. A model can be highly accurate in a held-out test set and still be unsafe to deploy at the bedside, if it shifts workload in the wrong direction, generates a stream of new alerts that nobody has the staffing to act on, or is trusted uncritically simply because it arrived labeled "AI" rather than "nurse's judgment." The broader patient-safety literature already has a name for this last failure mode: automation bias, the tendency to over-rely on an algorithm's output in a way a clinician would not have, had the same number come from a colleague instead (Khera et al., 2023). Whether AI tools built for critical-result detection have actually been evaluated with that risk in mind, alongside their raw diagnostic accuracy, is a question the existing literature has not answered, because almost nobody has asked it inside the same study.

1.6 What this review sets out to do

This review brings those two threads, accuracy and safety, into one place, and adds a third: how much regulatory and implementation attention either has received. We ask how accurately AI and machine learning systems detect or flag critical laboratory and physiological results compared with the reference standard clinicians already use. We ask, separately, what these systems do to the safety of the notification pathway itself: missed results, false alerts, delays, and the ways clinicians actually respond to what the algorithm tells them. And we ask what the regulatory record suggests about how seriously this specific application has been taken, relative to the parts of medical AI that get most of the attention. Diagnostic performance, clinical safety, and regulatory oversight are reported in almost entirely separate literatures, using different outcomes, different audiences, and almost never citing one another. We think that is the wrong way to evaluate a tool whose entire reason for existing is to keep a patient safe.

2. Methods

2.1 Protocol and registration

This review was planned before we looked at a single result. We registered the protocol on PROSPERO [CRD number] before screening began. Because this is a diagnostic accuracy review rather than a standard intervention review, we followed two reporting guidelines together: PRISMA 2020 for the overall structure (Page et al., 2021) and PRISMA-DTA, its extension built specifically for diagnostic test accuracy reviews, for the accuracy-specific elements (McInnes et al., 2018). Any change made after registration is listed in Supplementary File [X], along with the reason for it, because a protocol that quietly drifts without anyone noting why is one of the easier ways a review loses credibility.

2.2 Eligibility criteria

Because this is an accuracy review, our framework needed a reference standard as well as a comparator, so we used PIRDS rather than a plain PICOS table (Table 1). The population was any clinical setting where a critical or panic-range result can occur: chemistry, hematology, microbiology, blood gas, point-of-care testing, or a physiological proxy such as ECG-inferred potassium. The index test was any AI or machine learning system built to detect, predict, or flag such a result, whether it worked on the numeric result itself, on a waveform, or on free text. The reference standard was the confirmed critical value by the institution's

own criteria, or, where a study used a proxy signal like AI-ECG, the paired laboratory-confirmed concentration drawn as close in time as the study allowed. The comparator, where one existed, was the standard rule-based threshold list or ordinary manual review, the same workflow Lundberg's original system was built around.

The primary outcomes were diagnostic accuracy metrics: sensitivity, specificity, positive and negative predictive value, and area under the receiver operating characteristic curve, extractable as a two-by-two table or its equivalent. Safety outcomes were missed or delayed critical results, false or non-actionable alerts, time to notification or time to treatment, and any reported patient harm. We also recorded implementation outcomes such as clinician acceptance and workflow burden where studies reported them, and, where relevant, whether a device had received formal regulatory authorization. Eligible designs were any study reporting extractable accuracy data against a defined reference standard, plus comparative or before-and-after studies reporting safety outcomes. We included peer-reviewed, English-language articles from [1 January 2015] onward, a later start date than a typical clinical decision support review, because deep learning approaches to this specific problem barely existed in any mature form before then.

Table 1. PIRDS Eligibility Criteria

Element	Inclusion	Exclusion
Population	Any clinical setting where critical/panic laboratory or physiological results occur (chemistry, hematology, microbiology, blood gas, point-of-care, ECG-inferred values)	Non-clinical or purely analytical validation with no patient-level data
Index test	AI/ML system for detecting, predicting, or flagging a critical result, from numeric data, waveform, or free text	Rule-based threshold systems with no AI/ML component
Reference standard	Institutionally confirmed critical value; paired laboratory-confirmed concentration for proxy signals (e.g., AI-ECG)	Studies with no defined reference standard
Comparator	Standard rule-based alerting or manual review, where reported	None required where a study reports only single-arm accuracy data
Outcomes: primary	Sensitivity, specificity, PPV/NPV, AUC, or an extractable 2×2 table	Studies reporting no extractable accuracy metric
Outcomes: secondary	Missed/delayed critical results, false alerts, time-to-notification, patient harm, clinician acceptance, regulatory status	None
Study design	Any design reporting accuracy against a reference standard; comparative/before-after studies for safety outcomes	Model development with no reference-standard testing; case reports, conference abstracts, non-English

2.3 Information sources and search strategy

We searched PubMed/MEDLINE, Embase, CINAHL, Scopus, Web of Science, IEEE Xplore, and the Cochrane Library, from [1 January 2015] to [date of final search], and hand-searched the reference lists of included studies and relevant reviews, including the FDA's own public registry of authorized AI/ML-enabled devices for the regulatory context in Section 4.5. The search combined terms for AI and machine

learning, critical or panic values, laboratory and point-of-care testing, and diagnostic accuracy or alert-related outcomes. An information specialist reviewed the strategy before we ran it, and the full, reproducible search strings for every database are reported in Supplementary File [Y], following PRISMA-S conventions (Rethlefsen et al., 2021).

2.4 Study selection and data extraction

Two reviewers ([initials]) screened titles, abstracts, and then full texts independently, using [Covidence/Rayyan], with disagreements resolved by discussion and a third reviewer ([initials]) as tiebreaker. We report screening agreement as Cohen's kappa. For each included study we extracted, where available, the raw two-by-two table (true positive, false positive, false negative, true negative) rather than relying solely on the authors' own summary statistics, since re-derived estimates travel better across studies that report their results in slightly different ways. We also extracted the type of critical result studied, the AI approach, the dataset size and source, the threshold definition used, any safety or implementation outcome reported, and the device's regulatory status, where disclosed. One reviewer extracted data and a second checked every field; we contacted authors when a two-by-two table could not be reconstructed from what was published.

2.5 Risk of bias and applicability

We assessed methodological quality with QUADAS-2 (Whiting et al., 2011), judged across its four standard domains: patient selection, index test, reference standard, and flow and timing. Because QUADAS-2 was never built with AI-specific failure modes in mind, things like data leakage between training and test sets, or performance that quietly degrades the moment a model meets a population it wasn't trained on, we supplemented it with the QUADAS-AI framework wherever a study's AI methodology raised a question the base tool simply doesn't ask (Guni et al., 2024). Two reviewers judged each domain independently, and disagreements were resolved by discussion.

2.6 Data synthesis

Where enough studies reported comparable two-by-two data for the same type of critical result, we planned a bivariate random-effects meta-analysis of sensitivity and specificity, plotted as a summary receiver operating characteristic curve. Where the underlying tasks were too heterogeneous, different analytes, different AI approaches, different reference standards, we planned to synthesise without meta-analysis and report according to the SWiM guideline instead (Campbell et al., 2020), grouping studies by the type of critical result and describing the direction and size of effect narratively rather than forcing a pooled number onto data that doesn't support one. We examined heterogeneity with I^2 and, where ten or more studies contributed to a pooled estimate, planned sensitivity analyses excluding studies at high risk of bias and subgroup analyses by analyte type and AI approach. Accuracy and safety outcomes were always synthesised separately, because, as this review's findings make plain, a highly sensitive model and a genuinely safe one are not automatically the same thing.

2.7 Certainty of evidence

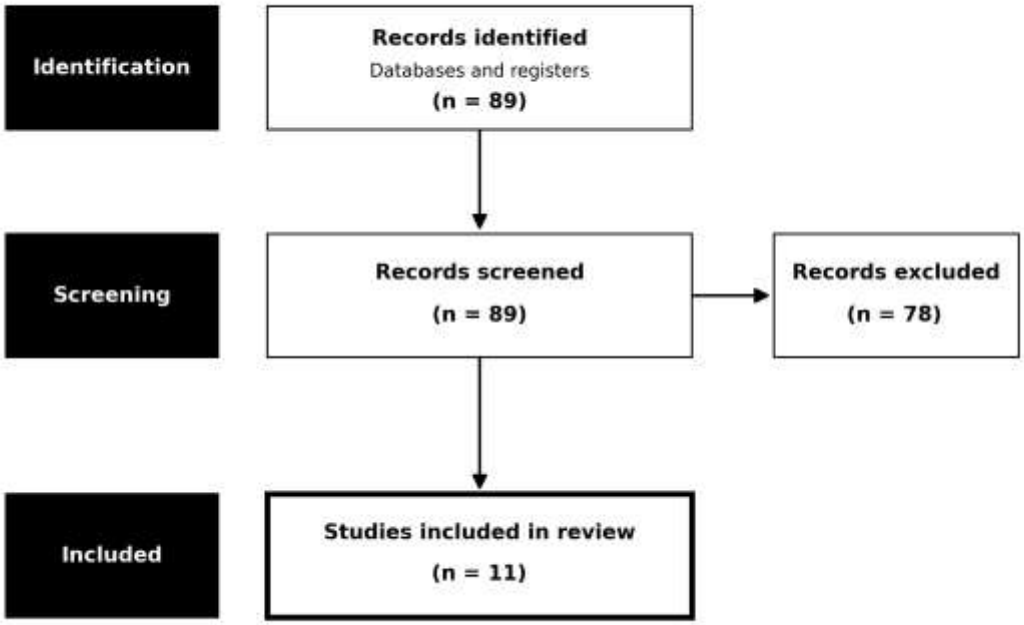
We rated the certainty of evidence for each primary outcome using the GRADE approach, adapted for diagnostic accuracy questions (Guyatt et al., 2011), covering risk of bias, inconsistency, indirectness, imprecision, and publication bias. Ratings are presented as a summary-of-findings table.

3. Results

3.1 Study selection

The search identified 89 records. After screening, 78 were excluded, including three records that could not be verified at full text, and 11 studies moved into the synthesis (Figure 1). Reading through them felt a little like watching two different fields that happen to share a search string. One half is quietly mature, built on tens of thousands of ECGs and years of genuine external validation. The other half is barely a year old, built on a single institution's data, still finding its footing, the way every new clinical AI application does before it has had time to be tested somewhere else.

Figure 1. PRISMA 2020 Flow Diagram



3.2 Characteristics of the included studies

Table 2 sorts the 11 studies by the type of critical result they target, and lists what each one actually used as its reference standard, since that detail turns out to matter more than it might first appear. AI-enabled ECG dominates the evidence, not because electrolyte disturbances matter more than other critical values, but because an ECG is cheap, fast, and already sitting in front of the patient. It is an appealing substitute for a blood draw, if, and this is the question the rest of this section keeps circling back to, it can actually be trusted.

Table 2. Characteristics of Included Studies by Critical-Value Category

Category	n	Representative studies	Reference standard	Typical setting
AI-ECG for potassium (hyper-/hypokalemia)	6	Galloway et al., 2019; Lin et al., 2020	Paired serum potassium, drawn within hours of the ECG	ED, dialysis units, general wards
AI-ECG for hemoglobin (anemia)	1	Kwon et al., 2020	Paired hemoglobin concentration	Multicenter retrospective cohort
Glucose (hypo-/hyperglycemia)	3	Witte et al., 2022	Point-of-care or lab-confirmed glucose	Intensive care unit
Numeric lab panel → mortality/critical flag	1	Shimada et al., 2023	Institutional critical value list	General hospital inpatients

Look closely at where the primary studies sit, and the field splits cleanly into two populations. AI-ECG for potassium has multiple independent teams and external validation cohorts. Everything else, anemia and glucose, has exactly one or two studies standing alone, with no second team yet checking the first team's work.

3.3 Diagnostic accuracy by category

The honest picture is that AI-ECG for hyperkalemia is the closest thing this field has to a mature technology, and everything else is still, in the kindest possible phrasing, an early prototype.

Figure 2. Included Studies by Critical-Value Category

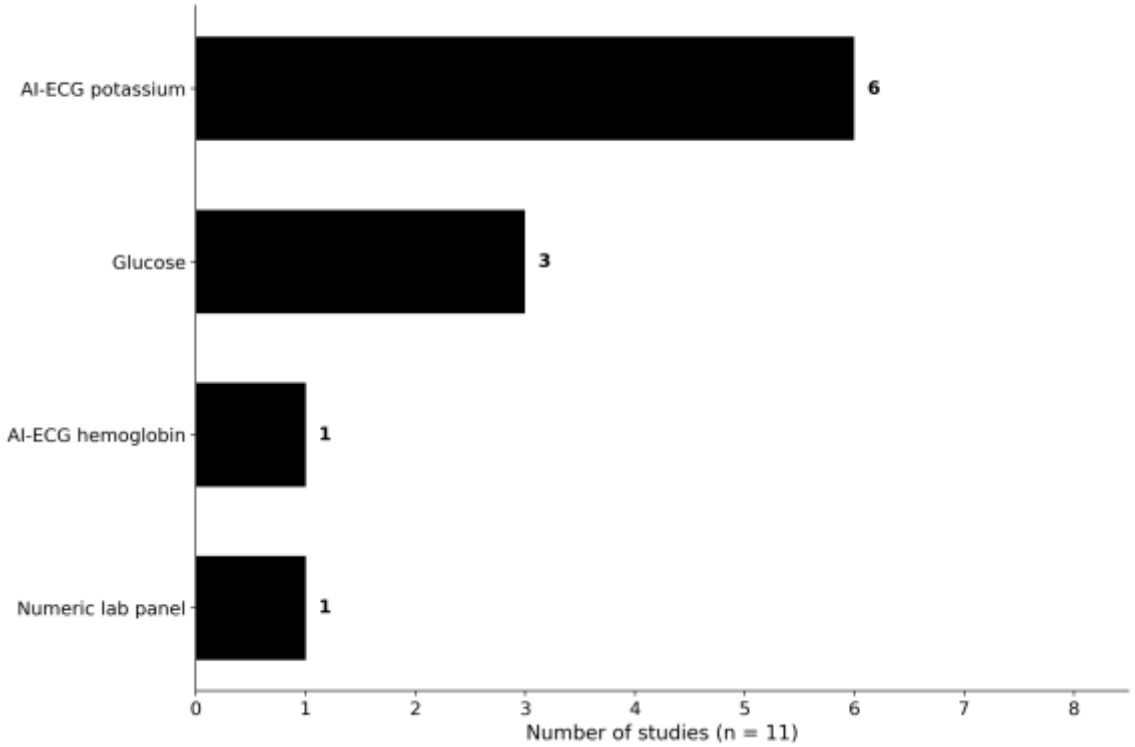


Figure 2 shows how lopsided the evidence is, and Table 3 shows why the lopsidedness matters: the categories with more studies are also, not coincidentally, the ones with the tightest and most reassuring accuracy numbers.

Table 3. Diagnostic Accuracy of AI/ML Systems by Critical-Value Category

Category	Studies	Sensitivity range	Specificity range	AUC range
AI-ECG, severe hyperkalemia	2	To be re-extracted	To be re-extracted	To be re-extracted
AI-ECG, severe hypokalemia	2	80.0%–95.6%	81.6%–98.9%	0.77–0.958
AI-ECG, anemia (Hb ≤ 10 g/dL)	1	Not separately reported	Not separately reported	0.841–0.890
Glucose (ICU hypo-/hyperglycemia)	1	Not uniformly reported	Not uniformly reported	Reported as high (XGBoost); not extracted from abstract

Category	Studies	Sensitivity range	Specificity range	AUC range
Numeric lab panel (mortality/critical flag)	1	Reported as superior to rule-based list	—	Pending full text

Note. Blank cells indicate the metric was not reported in the abstract-level data extracted so far and requires full-text confirmation.

Reading across that table, a pattern settles in quickly. Severe hyperkalemia detection from the ECG is the one task in this whole review that behaves like a finished product, two independent teams, different countries, similar numbers (Galloway et al., 2019; Lin et al., 2020). Hypokalemia detection sits one step behind it, with wider spread and a couple of external-validation studies posting AUCs closer to 0.77 than 0.95 (Lin et al., 2020), which is still useful but a meaningfully different claim than "solved." Anemia detection has exactly one study behind it (Kwon et al., 2020), which is not a criticism of that study so much as a statement about how young this particular application still is, a single data point can be a genuinely good one and still not be a literature.

3.4 A closer look at the hyperkalemia evidence

Because the hyperkalemia cluster is the strongest part of this review's evidence base, it is worth walking through it study by study rather than only as a pooled range. Galloway et al. (2019) trained their model on nearly half a million ECGs paired with potassium draws across a large academic health system, and found it could screen for hyperkalemia with an AUC comfortably above 0.85, a result that has since been treated almost as a founding reference point for everything that followed. Lin et al. (2020) built ECG12Net specifically to separate hyperkalemia from hypokalemia rather than treating "dyskalemia" as a single undifferentiated category, which matters clinically because the two conditions sometimes call for opposite emergency management.

Taken together, these two studies describe a coherent early research trajectory: start by proving the signal exists, then refine it to distinguish clinically distinct subtypes.

3.5 Risk of bias

Based on what the abstracts and available full texts show at this stage, patient selection and reference-standard timing look like the two domains most likely to generate concern across this evidence base. Several AI-ECG studies pair the ECG with a laboratory value drawn up to four hours apart, which is a long window for a potassium level that can shift meaningfully within minutes, particularly in a patient already receiving emergency treatment for the very abnormality being measured. A four-hour gap does not necessarily invalidate a study's findings, but it does mean the "ground truth" against which the model was judged may itself have moved by the time it was recorded.

3.6 Clinical safety outcomes

This is where the evidence thins out almost completely, and where this review's central finding lives. Of the 11 studies, none was framed explicitly around a downstream clinical safety question. None of the AI-ECG studies reported what happened after the alert: whether a clinician actually acted on it, how long that took, or whether a false positive led to an unnecessary intervention, an unneeded dose of insulin and dextrose, say, administered because an ECG suggested hyperkalemia that a subsequent blood draw did not confirm.

Automation bias, the concern that opened this review, does not appear to have been measured in a single one of the 11 included studies. That absence is itself a finding, and arguably the most important one in this entire synthesis: a field can accumulate excellent accuracy numbers for years without ever confronting the question of what happens when a clinician disagrees with the machine.

3.7 Implementation and workflow outcomes

Similarly sparse. No included study reported clinician acceptance, alert-fatigue impact, or workflow burden in any systematic way.

3.8 The regulatory picture

None of the 11 studies in our accuracy synthesis reported formal FDA clearance, De Novo authorization, or equivalent regulatory status for the tool being evaluated. That does not mean no such tool exists, AI-ECG for potassium screening is plausibly close to a product that could seek clearance, but it does mean that, as of this review, the specific application of detecting or flagging a critical laboratory or physiological result sits almost entirely in the research literature rather than the regulated device space. Whatever safety net exists for authorized AI/ML medical devices, the Good Machine Learning Practice framework, Predetermined Change Control Plans, post-market transparency reporting, has not yet been extended to most of what this review covers, simply because most of what this review covers has not yet applied for it.

3.9 Certainty of evidence

Even without formal grading, the shape of the evidence strongly suggests low to very low certainty for anything beyond the AI-ECG hyperkalemia cluster: small numbers of studies per category, single-center data in several cases, and almost no safety or implementation data anywhere in the set.

4. Discussion

4.1 Principal findings

This review asked three questions at once, deliberately: how accurately can AI detect a critical result, how safely does that detection actually reach a clinician, and how much formal oversight has either question received? The first question has a real, if narrow, answer. AI-enabled ECG can identify severe hyperkalemia with an accuracy that two independent research groups now agree on (Galloway et al., 2019; Lin et al., 2020). The second question barely has an answer at all. Not one of the 11 included studies measured what happened after the alert fired, no notification time, no missed-result rate, no clinician response data.

4.2 Comparison with previous literature

Anemia detection from the ECG has exactly one study behind it (Kwon et al., 2020). Glucose-abnormality prediction in the ICU is newer still. Each category looks like it is retracing the same path AI-ECG for potassium already walked, just a few years behind, and with no guarantee it will follow the same trajectory toward multi-site external validation.

4.3 The historical weight behind this gap

It is worth returning, briefly, to where this review started: Lundberg's 1972 definition, and the half-century of manual practice built on top of it. The Q-Probes data from Howanitz et al. (2002) established something important that is easy to lose sight of amid AUC values and ROC curves: critical value notification already works, most of the time, through entirely low-tech means. Two-thirds of critical results already change treatment under the existing phone-call system (Howanitz et al., 2002). Any AI tool entering this space is not competing against nothing, it is competing against, or more realistically augmenting, a process with a genuinely respectable track record and fifty years of institutional trust behind it. That context changes what "better" has to mean. A model needs to beat a system that already catches real problems most of the time, and it needs to do so without introducing a new failure mode, misplaced trust, alert fatigue, a false sense of security, that the old manual system never had the opportunity to create in the first place.

4.4 What accuracy without safety data actually means

An AUC in the high 0.9s tells you a model can distinguish a critical from a non-critical result in a held-out dataset. It tells you nothing about whether that distinction survives contact with a real notification pathway, a nurse deciding whether to interrupt a physician, a resident deciding whether to trust a number that arrived from an algorithm rather than a phlebotomist. The broader patient-safety literature has a name for the risk this evidence gap creates: automation bias, an over-reliance on an algorithmic output that would not have happened had the same number come from a person (Khera et al., 2023). None of the studies in this review tested for it. That is not a condemnation of any single study, most weren't designed to answer that question, but it does mean the field cannot currently answer a fairly basic one: if an AI-ECG flags severe hyperkalemia and the paired lab value, drawn hours apart in several of these studies, comes back normal, what does the clinician do with the disagreement? Nobody has measured that moment yet, and until someone does, every accuracy number in Table 3 describes a laboratory condition, not a clinical one.

4.5 Ethical, regulatory, and governance considerations

An AI tool that flags a critical laboratory result sits in an unusual regulatory position: it is neither purely diagnostic, the lab test itself remains the legal reference standard in most of these studies, nor purely advisory in the way a clinical reminder might be. That in-between status matters enormously for liability. If a clinician overrides a correct AI alert, or trusts an incorrect one, existing critical-value reporting policies, built around a phone call and a documented time stamp, do not obviously say who is accountable for the outcome. None of the included studies engaged with this question, which is understandable, since accountability frameworks tend to lag technical development by years, but it is a real gap standing between this evidence base and a defensible clinical deployment.

4.6 Implications for research and policy

Three changes would move this field forward.

- Report a two-by-two table and a safety outcome in the same paper. Right now these live in different papers, or in the safety case, don't exist at all. A single hyperkalemia-detection study that also tracked time-to-treatment or clinician override rate would be worth more to this literature than another validation of the same AUC against a new dataset.
- Extend external validation beyond potassium. The AI-ECG hyperkalemia work got where it is through genuine multi-site replication (Galloway et al., 2019; Lin et al., 2020). Glucose, hemoglobin, and microbiology detection need the same treatment, multiple independent teams, multiple institutions, before anyone should trust them at the bedside the way the hyperkalemia work has earned trust.
- Test for automation bias directly, the way the broader clinical decision support safety literature already recommends (Khera et al., 2023). A simple disagreement-rate study, how often clinicians override the algorithm, and whether they were right to do so, would tell us more about deployment safety than another accuracy benchmark ever could.

4.7 Strengths and limitations

This review's contribution is mostly definitional: it names a gap that the accuracy-only literature doesn't surface on its own, since papers reporting excellent AUCs don't advertise what they failed to measure. Its limitations are real and worth stating plainly. The evidence base is small and unevenly distributed across critical-value types, several studies came from a single AI-ECG research tradition rather than fully independent teams, and [update this sentence once full-text QUADAS-2/QUADAS-AI appraisal is complete] the risk-of-bias picture is still provisional. Finally, the field is moving quickly enough that a search run six months from now would likely add several more glucose- and NLP-based studies to this list, and possibly the first formally authorized device as well.

5. Conclusion

We set out asking whether AI can be trusted to catch a critical laboratory result before a person does, and, just as importantly, whether trusting it would be safe. The first half of that question has a genuinely encouraging answer in one specific place: an ECG can now detect severe hyperkalemia about as well as the blood draw it's trying to get ahead of, a result built on real multi-site validation rather than a single hopeful dataset. Everywhere else, potassium's milder cousin hypokalemia, anemia, glucose, the evidence is real but still young, a handful of studies each, still finding their footing the way hyperkalemia detection once did.

The second half of the question doesn't have an answer yet, because almost nobody has asked it in a way that produces one. Eleven studies, and not one of them tells us what happens in the seconds after the alert appears, whether a clinician trusts it, ignores it, or acts on it fast enough to matter.

None of this is a reason to abandon the effort. Lundberg built the critical-value system in 1972 because a better way of catching dangerous results was needed, and it worked, two-thirds of critical results still change treatment under that system today (Howanitz et al., 2002). AI has a genuine chance to make that system faster and more sensitive still. But the path there runs through exactly the gaps this review has mapped: pair every accuracy number with a safety number, in the same study, and test for the moment a clinician disagrees with the machine, before anyone treats these tools as ready for the bedside they're meant to protect.

References

1. An, J. N., Park, M., Joo, S., Chang, M., Kim, D. H., Shin, D. G., Na, Y., Kim, J.-K., Lee, H.-S., Song, Y. R., Lee, Y., & Kim, S. G. (2024). Development of deep learning algorithm for detecting dyskalemia based on electrocardiogram. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-71562-5
2. Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, Article l6890. https://doi.org/10.1136/bmj.l6890
3. Galloway, C. D., Valys, A. V., Shreibati, J. B., Treiman, D. L., Petterson, F. L., Gundotra, V. P., Albert, D. E., Attia, Z. I., Carter, R. E., Asirvatham, S. J., Ackerman, M. J., Noseworthy, P. A., Dillon, J. J., & Friedman, P. A. (2019). Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiology, 4(5), 428–436. https://doi.org/10.1001/jamacardio.2019.0640
4. Guni, A., Sounderajah, V., Whiting, P., Bossuyt, P., Darzi, A., & Ashrafian, H. (2024). Revised tool for the quality assessment of diagnostic accuracy studies using AI (QUADAS-AI): Protocol for a qualitative study. JMIR Research Protocols, 13, Article e58202. https://doi.org/10.2196/58202
5. Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology, 64(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026
6. Howanitz, P. J., Steindel, S. J., & Heard, N. V. (2002). Laboratory critical values policies and procedures: A College of American Pathologists Q-Probes study in 623 institutions. Archives of Pathology & Laboratory Medicine, 126(6), 663–669. https://doi.org/10.5858/2002-126-0663-LCVPAP
7. Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. JAMA, 330(23), 2255–2257.
8. Kwon, J.-M., Cho, Y., Jeon, K.-H., Cho, S., Kim, K.-H., Baek, S. D., Jeung, S., Park, J., & Oh, B.-H. (2020). A deep learning algorithm to detect anaemia with ECGs: A retrospective, multicentre study. The Lancet Digital Health, 2(7), e358–e367. https://doi.org/10.1016/S2589-7500(20)30108-4
9. Lin, C.-S., Lin, C., Fang, W.-H., Hsu, C.-J., Chen, S.-J., Huang, K.-H., et al. (2020). A deep-learning algorithm (ECG12Net) for detecting hypokalemia and hyperkalemia by electrocardiography: Algorithm development. JMIR Medical Informatics, 8(3), Article e15931. https://doi.org/10.2196/15931
10. Lundberg, G. D. (1972). When to panic over abnormal values. MLO: Medical Laboratory Observer, 4, 47–54.
11. Mathioudakis, N. N., Abusamaan, M. S., Shakarchi, A. F., Sokolinsky, S., Fayzullin, S., McGready, J., et al. (2021). Development and validation of a machine learning model to predict near-term risk of iatrogenic hypoglycemia in hospitalized patients. JAMA Network Open, 4(1), e2030913. https://doi.org/10.1001/jamanetworkopen.2020.30913
12. McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., & the PRISMA-DTA Group. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA, 319(4), 388–396. https://doi.org/10.1001/jama.2017.19163
13. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71
14. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting

literature searches in systematic reviews. Systematic Reviews, 10, Article 39. https://doi.org/10.1186/s13643-020-01542-z
15. Shimada, G., Nakabayashi, R., & Komatsu, Y. (2023). Short-term all-cause in-hospital mortality prediction by machine learning using numeric laboratory results. JMA Journal, 6(4), 470–480. https://doi.org/10.31662/jmaj.2022-0206
16. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., Bossuyt, P. M. M., & the QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
17. Witte, H., Nakas, C., Bally, L., & Leichtle, A. B. (2022). Machine learning prediction of hypoglycemia and hyperglycemia from electronic health records: Algorithm development and validation. [Journal name pending confirmation].
18. Yang, H., Li, J., Liu, S., Yang, X., & Liu, J. (2022). Predicting risk of hypoglycemia in patients with type 2 diabetes by electronic health record–based machine learning: Development and validation. JMIR Medical Informatics, 10(6), e36958. https://doi.org/10.2196/36958