International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

# Advancing Healthcare through Artificial Intelligence: The Role of Association Rule Mining in Clinical Decision Support and Healthcare Analytics

**Authors: Geofrey Nyabuto; Charles Kibet Ng'etich; Edwin Seno; George Kihara Mburu; Marion Jeptoo; Joyadams Munene; Muriithi Alex Karani; John kimani Muragu; Nyairo Charles Magati**

**Affiliation**: Open University of Kenya

**Abstract- Association rule mining (ARM) is a data mining approach used to discover frequent co-occurrence patterns and conditional relationships in large datasets. In healthcare, ARM has been applied to electronic health records, claims databases, laboratory data, prescription data, disease registries, and public health datasets to reveal clinically meaningful patterns that may support diagnosis, medication safety, risk stratification, and service planning. Objective: This review synthesizes how ARM has been applied in healthcare, focusing on methods, clinical application areas, implementation challenges, and future research directions. A systematic review design guided by PRISMA 2020 was used to structure the manuscript. Literature was organized around peer-reviewed ARM studies in healthcare, including clinical decision support, diagnostic test ordering, disease-medication association mining, adverse drug reaction signal detection, risk factor discovery, hospital readmission analysis, privacy-preserving mining, and emerging causal or hybrid ARM approaches. The literature shows that Apriori remains the most frequently used ARM algorithm, although FP-Growth, weighted Apriori, class association rules, negative association mining, privacy-preserving ARM, and causal irredundant ARM are increasingly used to address computational, interpretability, privacy, and clinical validity limitations. ARM is valuable because it produces transparent IF-THEN rules that clinicians can inspect, but uncontrolled rule generation, weak validation, data quality limitations, and spurious associations remain major barriers. ARM has clear potential in healthcare knowledge discovery and decision support, particularly where interpretability is required. Future research should prioritize external validation, clinician-centered rule evaluation, integration with electronic medical records, explainable hybrid models, privacy-preserving analytics, and evidence from low- and middle-income healthcare settings.**

**Keywords: Association rule mining; Apriori algorithm; healthcare data mining; clinical decision support; electronic health records; systematic review; adverse drug reaction; health informatics.**

## I. INTRODUCTION

Healthcare systems generate large volumes of heterogeneous data through electronic health records, laboratory information systems, pharmacy systems, claims databases, disease registries, public health reporting platforms, imaging repositories, and patient-generated digital tools (Javeedullah, 2025). These datasets are valuable not only for reporting and operational monitoring but also for discovering patterns that can improve clinical decision-making, patient safety, disease prevention, and health system efficiency. However, much of the knowledge contained in these datasets remains hidden because relationships among diagnoses, medications, laboratory results, risk factors, and outcomes are often complex, multidimensional, and difficult to identify using manual review alone (Holmes et al., 2021).

Association rule mining (ARM) is one of the core methods within knowledge discovery in databases. The classical idea of ARM was popularized by Agrawal, Imielinski, and Swami (1993), who introduced association rules to discover relationships among items in large transactional databases. In simple terms, ARM searches for rules of the form X -> Y, meaning that when itemset X occurs, itemset Y is likely to occur as well. In healthcare, X and Y may represent diagnoses, symptoms, medications, laboratory abnormalities, procedures, demographic characteristics, behavioral risk factors, or outcomes. A rule such as $\{\text{diabetes}, \text{hypertension}\} \rightarrow \{\text{renal function test required}\}$ may indicate a common clinical practice pattern, whereas $\{\text{drug A}, \text{drug B}\} \rightarrow \{\text{adverse event}\}$ may suggest a pharmacovigilance signal that requires further review.

The appeal of ARM in healthcare is its interpretability. Unlike many black-box machine learning models, association rules can be expressed in plain language and reviewed by clinicians (Ilma et al., 2023). Rule metrics such as support, confidence, lift, conviction, leverage, and chi-square provide ways of evaluating the frequency, conditional probability, and strength of discovered relationships. Support indicates how often a rule occurs in the dataset; confidence estimates how often the consequent occurs when the antecedent occurs; and lift compares the observed co-occurrence with what would be expected under independence. This interpretability makes ARM attractive for clinical decision support, medication safety, guideline auditing, and public health surveillance.

Earlier and contemporary reviews show that ARM has been applied across many health informatics problems, including disease prediction, comorbidity discovery, adverse drug reaction detection, medical knowledge discovery, lifestyle risk behavior analysis, and clinical decision support (Altaf et al., 2017; Sarıyer & Taşar, 2019). Recent studies continue to extend ARM beyond basic Apriori mining toward weighted, causal, negative, hybrid, and privacy-preserving methods. For example, ARM has been applied to identify medication-problem and laboratory-problem associations in electronic health records (Wright et al., 2010), support real-time ICU decision support (Cheng et al., 2013), discover diagnosis-laboratory test relationships in emergency departments (Sarıyer & Öcal Taşar, 2020), reduce pseudo-associations in diagnosis-medication mining (Wang et al., 2021), and detect adverse drug reaction signals using large administrative claims data (Yamamoto et al., 2023).

Despite its usefulness, ARM faces important challenges in healthcare. Healthcare datasets contain missing values, coding inconsistencies, duplicated records, irregular time intervals, changing clinical guidelines, and confounding relationships (Bayley et al., 2013). ARM can generate large numbers of rules, many of which may be statistically interesting but clinically irrelevant. High confidence may reflect the high prevalence of common outcomes rather than a meaningful relationship. Lift may overstate rare associations. In addition, rules discovered from retrospective datasets may not generalize across facilities, regions, populations, or care models. These challenges require careful preprocessing, appropriate threshold selection, expert validation, external validation, and transparent reporting. The objective of this article was to systematically review the application of ARM in healthcare with emphasis on methods, clinical applications, challenges, and future directions.

## II. METHODS

This review was structured according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidance, which recommends transparent reporting of the review rationale, eligibility criteria, information sources, search strategy, selection process, data items, synthesis methods, and results (Page et al., 2021a, 2021b). The review question was: How has association rule mining been applied in healthcare, and what methods, clinical applications, challenges, and future research directions are evident in the literature?

The conceptual population of interest was healthcare-related datasets, including electronic health records, administrative claims, pharmacy records, laboratory records, hospital information systems, intensive care data, public health surveys, and disease-specific registries. The intervention or method of interest was association rule mining, including Apriori, FP-Growth, Eclat, weighted Apriori, class association rules, negative association rules, causal association rules, privacy-preserving association rule mining, and hybrid machine learning models that incorporate association rules. Outcomes of interest included discovered clinical rules, decision support applications, medication safety signals, diagnostic patterns, disease risk factors, readmission patterns, comorbidity structures, interpretability, validation approaches, and methodological limitations.

Search terms for a full systematic review combined method and healthcare context. Recommended Boolean strings included: ("association rule mining" OR "association rules" OR Apriori OR "frequent pattern mining" OR "FP-Growth" OR "class association rules") AND (healthcare OR medical OR clinical OR hospital OR patient OR "electronic health records" OR EHR OR EMR OR pharmacy OR diagnosis OR medication OR "adverse drug reaction" OR comorbidity). The most relevant databases for searching included PubMed, Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar for supplementary searching. Searches were limited to peer-reviewed journal articles and conference papers, with the final date range selected as between 2010 to 2026, because this captures early EHR-based ARM studies and

recent developments in privacy-preserving, hybrid, and causal ARM.

Inclusion criteria was: (1) the study applied ARM or a closely related frequent pattern mining method to healthcare or biomedical data; (2) the study reported rule mining methods or rule metrics such as support, confidence, lift, conviction, interest, or related measures; (3) the study used empirical healthcare data or clinically relevant simulated data; (4) the study reported an application, methodological contribution, or validation relevant to healthcare decision-making; and (5) the article was peer-reviewed and available in English. Exclusion criteria should be: (1) non-healthcare applications; (2) papers that merely mention ARM without applying it; (3) commentaries, editorials, and opinion papers without empirical or methodological contribution; (4) non-peer-reviewed web materials; and (5) duplicate records.

During the review of the journal articles and conference proceedings, at least 2 reviewers independently screened the titles, abstracts, and full texts. Disagreements were resolved through discussions and third reviewer. Data extraction captured author, year, country or dataset setting, healthcare domain, dataset type, algorithm, metrics, validation approach, main findings, limitations, and contribution. Because ARM studies are heterogeneous in algorithms, datasets, outcomes, and evaluation metrics, narrative thematic synthesis is more appropriate than meta-analysis.

The synthesis in this article was organized around methodological trends and application domains. The goal was not to pool effect sizes but to clarify how ARM has been used, what it contributes, and where methodological weaknesses remain.

**Table 1. Recommended search strategy and eligibility structure for the systematic review.**

| Component | Recommended content |
|---|---|
| Review framework | PRISMA 2020 reporting guidance; narrative synthesis due to heterogeneity of algorithms and outcomes. |
| Databases | PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Google Scholar. |
| Core search terms | "Association rule mining", "association rules", Apriori, "frequent pattern mining", "FP-Growth", healthcare, clinical, EHR, EMR, diagnosis, medication, adverse drug reaction, comorbidity. |
| Inclusion criteria | Empirical or methodological ARM studies using healthcare data; reported algorithms or rule metrics; peer-reviewed; English language. |
| Exclusion criteria | Non-healthcare studies, opinion papers, duplicate records, studies that mention but do not apply ARM, and unavailable full texts. |
| Data extraction | Author, year, country/dataset setting, healthcare domain, data type, algorithm, metrics, validation method, findings, limitations. |
| Synthesis approach | Thematic synthesis by methods and clinical application area; no statistical pooling unless comparable outcomes are available. |
| PRISMA flow | Insert final numbers after database searching, duplicate removal, title/abstract screening, full-text review, and final inclusion. |

## III. RESULTS AND SYNTHESIS OF THE LITERATURE

The literature reviewed indicates that ARM in healthcare can be grouped into eight major application areas: clinical decision support, diagnostic test ordering, medication-problem and diagnosis-medication association mining, adverse drug reaction signal detection, disease risk factor discovery, hospital readmission and utilization analysis, lifestyle and public health pattern discovery, and privacy-preserving or advanced methodological extensions. Across these areas, ARM is valued because it transforms large clinical datasets into interpretable rules that can be reviewed by clinicians, informaticians, pharmacists, and managers.

The Apriori algorithm remains the most visible method in applied healthcare studies. It is easy to understand, available in common analytical software, and produces rules that can be explained to clinical

stakeholders. However, Apriori requires repeated scanning of data and can become computationally expensive in large, high-dimensional datasets. FP-Growth is used where candidate generation becomes costly, while weighted Apriori and modified mining strategies are used where variables differ in clinical importance. More recent studies show movement toward class association rules, causal ARM, negative ARM, and privacy-preserving ARM to address limitations of simple frequent co-occurrence mining.

Electronic health records and claims databases dominated the evidence base. This was expected because ARM requires transactional or binary representations of co-occurrence. Healthcare data are commonly transformed into patient-item matrices, where items may include diagnoses, medications, laboratory orders, symptoms, procedures, demographic groups, or risk factors. Some studies used large routine databases: Wright et al. (2010) analyzed structured EHR data from 100,000 patients to identify medication-problem and laboratory-problem associations; Wang et al. (2021) analyzed over one billion prescriptions from claims data to reduce diagnosis-medication pseudo-associations; and Narindrarangkura et al. (2023) applied ARM to multi-center real-world diabetes data to identify risk patterns associated with suicide attempts.

Validation practices varied substantially. Some studies used expert review, as in emergency department diagnostic test rule validation by practitioners (Sarıyer & Taşar, 2020). Others compared rules with known drug references, laboratory references, or clinical standards, as in Wright et al. (2010). Some papers assessed predictive performance using accuracy, sensitivity, specificity, F1 score, or related metrics when association rules were used as classifiers, as in differentiated thyroid cancer recurrence prediction (Firat Atay et al., 2024). However, many ARM studies still lack external validation across sites, prospective testing, and evidence that discovered rules improve clinical outcomes.

The evidence also suggests that ARM is shifting from descriptive knowledge discovery toward decision support and explainable prediction. Early ARM studies often focused on discovering interesting patterns; recent work increasingly integrates ARM into clinical decision support, hybrid machine learning workflows, pharmacovigilance, and privacy-preserving analytics. This shift is important because healthcare decision-makers require not only interesting patterns but also validated, actionable, ethically acceptable, and clinically interpretable rules.

**Table 2. Representative studies showing how ARM has been applied in healthcare.**

| Study | Healthcare area | Data type | ARM method | Main contribution |
|---|---|---|---|---|
| Wright et al. (2010) | Clinical knowledge discovery | Structured EHR data | Association rules with multiple metrics | Identified medication-problem and laboratory-problem associations useful for problem-list improvement. |
| Wright et al. (2013) | Medication-problem validation | EHR medication and problem lists | Association rule mining | Validated ARM for inferring medication-problem associations. |
| Cheng et al. (2013) | ICU decision support | ICU clinical data | Real-time ARM interface | Developed icuARM to support interactive |

| Study | Healthcare area | Data type | ARM method | Main contribution |
|---|---|---|---|---|
| | | | | ICU rule mining. |
| Sarıyer & Öcal Taşar (2020) | Emergency diagnostics | ED diagnosis and test data | Apriori; positive and negative rules | Mapped diagnosis categories to laboratory diagnostic test requirements and obtained expert validation. |
| Miswan et al. (2021) | Hospital readmission | Hospital admission/readmission data | Apriori | Discovered demographic and readmission-related patterns. |
| Wang et al. (2021) | Medication safety | Claims prescriptions and diagnoses | Modified lift-based ARM | Reduced diagnosis-medication pseudo-associations in large claims data. |
| Wei et al. (2021) | Pharmacovigilance | Adverse drug reaction data | Association rule analysis | Used ARM metrics to mine adverse drug reaction signals. |
| Odu et al. (2022) | Tuberculosis decision support | TB-related digital health data | FP-Growth | Generated recurring diagnostic and co-occurrence rules for drug-resistant TB decision support. |
| Yamamoto et al. (2023) | Adverse drug reaction detection | Administrative claims data | ARM for signal detection | Used large claims data to detect ADR signals earlier. |

| Study | Healthcare area | Data type | ARM method | Main contribution |
|---|---|---|---|---|
| Narindrarangkura et al. (2023) | Diabetes and mental health risk | Multi-center real-world EHR data | Apriori after feature selection | Identified explainable risk patterns for suicide attempts among people with diabetes. |
| Guillamet et al. (2023) | Patient trajectories | Primary care visits | Causal irredundant ARM | Reduced redundant rules and assessed potential causal paths while controlling confounders. |
| Budarajju & Jammalamadaka (2024) | Medical negative associations | Medical database examples | Negative ARM with closed/maximal itemsets | Reduced rule volume and highlighted importance of negative associations in medical decision support. |
| Firat Atay et al. (2024) | Cancer recurrence prediction | Differentiated thyroid cancer data | Class association rules | Combined ARM and classification to build interpretable recurrence prediction models. |
| Domadiya & Rao (2019, 2021) | Privacy-preserving healthcare mining | Distributed/partitioned healthcare data | Privacy-preserving ARM | Addressed privacy concerns in distributed ARM over healthcare data. |

# IV. CLINICAL APPLICATIONS OF ASSOCIATION RULE MINING IN HEALTHCARE

### 4.1 Clinical decision support and knowledge discovery

Clinical decision support is one of the most important application areas of ARM because health professionals often need transparent evidence that can be inspected and discussed. ARM can convert routine clinical data into rules that suggest potential diagnoses, highlight missing documentation, recommend laboratory follow-up, identify common co-prescription patterns, or flag potential safety issues. Wright et al. (2010) demonstrated that ARM can identify clinically accurate associations between medications, laboratory results, and problems using structured electronic health record data. This type of work is particularly relevant for improving problem lists, which are often incomplete and can affect quality measurement, decision support, and continuity of care.

The icuARM system developed by Cheng et al. (2013) illustrates how ARM can be embedded into a clinical decision support interface. Intensive care units generate high-frequency, complex patient data, and clinicians must make timely decisions under uncertainty. An interactive ARM tool allows clinicians to specify clinical scenarios and retrieve rules that describe similar historical patterns. The strength of such systems is not that they replace clinical judgment but that they present interpretable evidence derived from comparable patient records. However, decision support in ICU highlights that rules must be current, validated, and clinically meaningful to avoid cognitive overload or unsafe recommendations.

### 4.2 Diagnostic test ordering and emergency department resource use

ARM has been applied to diagnostic test ordering, especially in emergency departments where patient flow, cost, and turnaround time are major concerns. Sarıyer and Öcal Taşar (2020) used Apriori to examine associations between diagnosis categories and laboratory diagnostic test requirements. Their study is important because it included both positive and negative rules and involved expert validation. Positive rules may indicate diagnoses for which particular tests are commonly required, while negative rules may indicate diagnoses where certain tests are usually unnecessary. In overstretched emergency departments, such rules may support more efficient resource use, reduce unnecessary testing, and guide laboratory capacity planning.

The relevance of ARM in this area is strongest when rules are interpreted as decision aids rather than rigid protocols. A rule indicating that a test is rarely ordered for a diagnosis category should not prevent clinicians from ordering that test when clinically justified. Instead, ARM can help identify overused tests, common ordering bundles, and potential areas for guideline review. Future research in this area should evaluate whether ARM-based recommendations reduce cost and waiting time without increasing missed diagnoses.

### 4.3 Medication, diagnosis, and prescribing pattern analysis

Medication-related ARM studies show strong practical value because prescriptions, diagnoses, and pharmacy transactions are naturally suited to transactional mining. Wang et al. (2021) used diagnosis-medication association mining to identify real-world disease-medication profiles while addressing pseudo-associations. This is a crucial problem where common comorbidities can create misleading associations. For example, if many patients with hypertension also have diabetes, a medication for diabetes may appear associated with hypertension unless the algorithm accounts for confounding prescription patterns. The study showed that modified mining strategies can improve the clinical credibility of mined associations.

Medication-problem association mining has also been used to infer missing clinical problems from prescriptions and laboratory results (Wright et al., 2010; Wright et al., 2013). Such work is relevant to electronic medical record quality because incomplete diagnosis lists can weaken clinical decision support and reporting. However, medication-problem rules must be interpreted carefully. Medication may be used for multiple indications, and off-label use, prophylaxis, and comorbid treatment can complicate the relationship between drug and diagnosis. Expert validation and linkage with drug reference standards remain necessary.

### 4.4 Adverse drug reaction and pharmacovigilance signal detection

Adverse drug reaction detection is another important area where ARM can reveal drug-event patterns. Wei et al. (2021) applied association rule analysis to mine adverse drug reaction signals, while Yamamoto et al.

(2023) used ARM with large-scale administrative claims data for earlier detection of adverse drug reaction signals. Pharmacovigilance data are challenging because adverse events may be rare, underreported, delayed, or confounded by underlying disease severity. ARM can help by scanning large datasets for non-obvious drug-event combinations, but identified rules should be treated as signals for investigation rather than proof of causality.

The role of lift and related interestingness measures is particularly important in pharmacovigilance. A high-confidence rule may simply reflect a common adverse event, whereas lift can highlight event frequencies greater than expected by chance. However, rare but serious adverse events may have low support and still be clinically important. Therefore, pharmacovigilance ARM should combine statistical thresholds with clinical severity, temporal plausibility, biological plausibility, and review by pharmacologists or clinicians.

### 4.5 Disease risk factor discovery and comorbidity analysis

ARM has been used to discover risk factor patterns for chronic diseases and comorbidities. Nahar et al. (2013) applied ARM to identify factors contributing to heart disease in males and females. Ramezankhani et al. (2015) applied ARM to extract risk patterns for type 2 diabetes using Tehran Lipid and Glucose Study data. Park et al. (2014) used ARM to understand clustering of lifestyle risk behaviors among Korean adults. These studies demonstrate that ARM can identify combinations of factors rather than isolated predictors. This is important in public health because risk behaviors and clinical risk factors often cluster in ways that require integrated interventions.

More recent studies show the continued value of ARM for complex risk discovery. Narindrarangkura et al. (2023) used ARM to uncover links among race, glycemic control, lipid profiles, and suicide attempts in individuals with diabetes. This type of study demonstrates ARM's usefulness in revealing interpretable high-risk subgroups from large multi-center datasets. Nevertheless, risk pattern discovery must avoid overinterpretation. Association does not imply causation and discovered risk combinations may be influenced by data availability, service utilization, coding practices, and social determinants not fully captured in structured data.

### 4.6 Hospital readmission and healthcare utilization

Hospital readmission is a major quality, cost, and performance indicator. Miswan et al. (2021) applied ARM to hospital readmission data and found associations among readmission length and demographic variables. ARM is useful in this area because readmission is influenced by multiple interacting factors, including age, comorbidity, admission type, discharge planning, medication burden, socioeconomic context, and service availability. Rule-based output can help managers and clinicians identify patient subgroups that may require stronger transition-of-care interventions.

However, readmission-related ARM requires careful design. A rule may identify patients who are frequently readmitted but may not explain preventability. Some re-admissions are clinically necessary, while others reflect gaps in discharge planning, outpatient follow-up, medication reconciliation, or social support. Therefore, ARM should be combined with clinical review and, where possible, intervention studies that test whether rule-guided care reduces avoidable readmissions.

### 4.7 Oncology and disease-specific predictive modelling

ARM is increasingly integrated into disease-specific predictive modelling. Firat Atay et al. (2024) combined association rule mining with classification algorithms to predict differentiated thyroid cancer recurrence. Their study showed how class association rules supported interpretable prediction: instead of producing only a probability score, the model presented rule structures linking clinicopathological variables with recurrence risk. This is highly relevant in precision medicine, where clinicians often require both predictive accuracy and explanation.

Hybrid models may represent a promising future direction because ARM alone is often descriptive, while classification models can estimate predictive performance. Combining ARM with logistic regression, random forests, gradient boosting, or neural networks can improve prediction while maintaining interpretability if rules are used for explanation, feature engineering, subgroup discovery, or post-hoc clinical interpretation.

### 4.8 Advanced, causal, negative, and privacy-preserving ARM

Recent methodological studies attempt to solve persistent ARM problems. Guillamet et al. (2023) developed CauRuler, a causal irredundant association rule miner for complex patient trajectory modelling. This is important because conventional ARM often discovers redundant and spurious associations. By reducing redundancy and controlling confounding variables, causal-oriented ARM can move closer to clinically meaningful trajectory analysis. Similarly, Budaraju and Jammalamadaka (2024) emphasized negative associations in medical databases, showing that the absence of co-occurrence or contraindicated combinations may be as important as positive co-occurrence. Negative association rules can be relevant in drug conflict detection, treatment incompatibility, and unusual clinical patterns.

Privacy-preserving ARM is another critical direction because healthcare data are sensitive and often distributed across institutions. Domadiya and Rao (2019) addressed privacy-preserving distributed ARM for vertically partitioned healthcare data, while later work extended secure approaches for distributed healthcare data mining. These methods are particularly relevant where hospitals cannot share raw patient data but need collaborative learning across sites. Privacy-preserving and federated ARM may become increasingly important for multi-site learning, national digital health platforms, and cross-border research.

**Table 3. Thematic synthesis of ARM contributions and limitations in healthcare.**

| Theme | Healthcare value | Typical rule examples | Key limitation |
|---|---|---|---|
| Decision support | Supports interpretable recommendations and documentation review. | {medication, lab a → {possible diag | Rules may be outdated or clinically irrelevant without validation. |
| Diagnostic tests | Improves understanding of test-ordering patterns and resource use. | {diagnosis catego → {test bundle} | Rare but necessary tests may be wrongly interpreted as unnecessary. |
| Medication safety | Detects prescribing patterns, pseudo-associations, and drug-event signals. | {drug combinatio → {adverse even | Confounding by indication and comorbidity can distort associations. |
| Risk factor discovery | Identifies high-risk subgroups using combinations of variables. | {risk factors} → {disease outco | Associations are not causal and may reflect coding bias. |
| Hospital utilization | Identifies subgroups linked to readmission or resource use. | {age group, como → {readmission | Readmission preventability is not directly established. |
| Privacy-preserving ARM | Enables multi-site learning without raw data sharing. | {distributed frequ → {shared rules} | Complex implementation and trust requirements. |
| Causal/negative ARM | Reduces redundancy and explores | {condition A} → {not medicatio | Requires careful assump |

| Theme | Healthcare value | Typical rule examples | Key limitation |
|---|---|---|---|
| | causal or absence-based patterns. | | tions and clinical interpretation. |

## V. APPLICATION OF ASSOCIATION RULE MINING IN KENYA'S HEALTHCARE SYSTEM

Association Rule Mining (ARM) is still a relatively new and underutilized data mining approach in Kenya's healthcare system. Although Kenya has made strong progress in digitizing health information through systems such as KenyaEMR, OpenMRS-based platforms, and DHIS2/KHIS, the routine use of ARM for clinical decision support, disease surveillance, patient risk profiling, and health-system planning remains limited. Most Kenyan health facilities and county health teams still rely mainly on descriptive dashboards, indicator reports, and routine data review meetings rather than advanced pattern-discovery techniques.

In 2014, Kang'ethe and Wagacha applied the Apriori algorithm to mine diagnosis patterns from electronic medical records. Their study demonstrated how association rules can be used to uncover relationships among patient diagnoses across multiple clinical encounters. Although the dataset used was not drawn directly from Kenyan health facilities, the study was important because it was conducted by Kenyan researchers and showed how EMR data could be transformed into clinically meaningful diagnosis patterns. The study found that ARM could confirm already known disease relationships while also revealing less obvious associations that may require further clinical investigation (Kang'ethe & Wagacha, 2014). This is important for Kenya because EMR systems such as KenyaEMR and other digital platforms are now common in HIV care and other health programs. As more patient-level data becomes available, ARM can support clinicians and health managers to identify comorbidities, treatment patterns, and high-risk patient groups earlier.

ARM has also been explored in cancer-related decision support in Kenya. Gatobu (2025) reported a hybrid breast cancer prediction model that incorporated association rule mining techniques, including Apriori and FP-Growth, together with classification approaches. This is important because breast cancer remains a major public health concern, and many patients in underserved communities are diagnosed late. ARM can contribute to cancer care by identifying combinations of risk factors, clinical symptoms, screening history, and demographic characteristics that frequently occur among patients with suspected or confirmed breast cancer. Unlike some machine learning models that only produce a prediction score, association rules can help explain why a patient may be considered high risk. This makes ARM useful in settings where clinicians and health workers need models that are not only accurate but also understandable and actionable.

Overall, ARM has not yet become a fully routine clinical decision-support tool in Kenya, but the direction is promising. Kenya already has the key ingredients needed for ARM adoption i.e., electronic medical records, KHIS/DHIS2 data, HIV program databases, cancer screening data, nutrition and consumer datasets, and growing local expertise in data science.

## VI. CHALLENGES AND LIMITATIONS

The first major challenge is data quality. ARM is highly sensitive to how data are coded, cleaned, and transformed. Missing diagnoses, incomplete medication lists, inconsistent laboratory units, duplicate patient records, and changes in coding systems can produce misleading rules. In electronic health records, absence of documentation does not always mean absence of disease, symptoms, or treatment. Therefore, preprocessing decisions such as binarization, discretization of continuous variables, handling of missing values, and grouping of codes into clinical categories should be reported clearly.

The second challenge is rule explosion. Healthcare datasets may contain thousands of diagnoses, medications, laboratory tests, procedures, and demographic variables. Even modest support and confidence thresholds can generate thousands or millions of candidate rules. Clinicians cannot review very large rulesets. This creates a need for pruning, ranking, grouping, and visualization. Interestingness measures should be selected carefully, and rule filtering should incorporate clinical relevance, actionability, novelty, and potential harm.

The third challenge is spurious association. ARM discovers co-occurrence, not causality. A strong rule may arise because two conditions are common, because a medication is prescribed for a comorbid disease, because a laboratory test is routinely ordered for administrative reasons, or because coding practices differ across facilities. The diagnosis-medication pseudo-association problem described by Wang et al. (2021) is a clear example of this issue. Confounding, temporal ambiguity, and institutional practice patterns can all produce rules that appear meaningful but are not clinically valid.

The fourth challenge is threshold selection. Minimum support, confidence, and lift thresholds strongly affect the rules discovered. High thresholds may miss rare but important clinical events, while low thresholds may produce too many weak rules. This is especially problematic in adverse drug reaction detection, rare disease research, and specialized oncology outcomes. Thresholds should be justified through sensitivity analysis, expert consultation, or validation against known standards.

The fifth challenge is validation. Many ARM studies validate rules internally but do not test whether the rules generalize to new hospitals, new time periods, or different populations. Clinical adoption requires more than statistical association; it requires evidence that rules are accurate, interpretable, actionable, and safe. Prospective validation and clinical workflow evaluation remain limited in the literature.

The sixth challenge is privacy and governance. ARM often requires access to detailed patient-level data. Multi-site ARM may be limited by legal, ethical, and institutional restrictions on sharing identifiable or sensitive data. Privacy-preserving ARM methods are promising, but they require technical capacity, governance agreements, and trust across participating institutions. Especially is Kenya, there exist Data Protection Act (DPA) that guides how personal data should be handles especially the sensitive patient data.

## VII. RESEARCH GAPS AND FUTURE DIRECTIONS

Future studies should prioritize clinically validated ARM. Many studies demonstrate that rules can be generated, but few show that rules improve care, reduce harm, lower cost, or improve patient outcomes. Stronger research designs should include external validation, temporal validation, expert adjudication, and prospective testing in real clinical workflows.

A second future direction is integration of ARM into electronic medical record systems. ARM outputs are most useful when they are embedded into the workflow at the point of care, pharmacy review, laboratory ordering, discharge planning, or public health surveillance. However, integration must avoid alert fatigue. ARM-based decision support should provide concise, high-value, explainable recommendations and should allow clinicians to review the evidence behind each rule.

A third direction is hybrid explainable modelling. ARM can be combined with machine learning models to support feature engineering, subgroup discovery, associative classification, and explanation of predictive outputs. For example, a random forest or gradient boosting model may provide high predictive accuracy, while ARM may provide interpretable subgroup rules that help clinicians understand why certain patients are high risk. Such hybrid approaches are especially important in precision medicine and chronic disease management.

A fourth direction is causal and temporal ARM. Many healthcare events occur over time, and simple co-occurrence rules may ignore the sequence of symptoms, diagnoses, treatment, and outcomes. Temporal ARM, sequential pattern mining, and causal rule mining can help identify whether one event tends to precede another and whether an association remains robust after controlling for confounders. Patient trajectory modelling, as illustrated by causal irredundant ARM, is particularly promising for chronic disease progression and multimorbidity research.

The fifth direction is privacy-preserving and federated ARM. As healthcare systems digitize, there is increasing demand for learning across institutions without centralizing sensitive patient data. Federated and privacy-preserving ARM can support collaborative knowledge discovery while respecting data protection requirements. This is especially relevant for national digital health ecosystems, rare disease research, and multi-county or multi-facility health analytics.

A sixth direction is research in low- and middle-income countries. Much of the ARM literature is based on high-income settings with mature electronic health record infrastructure. There is need for evidence from settings where data quality, interoperability, human resources, and infrastructure constraints differ. ARM could be valuable in HIV, tuberculosis, malaria,

maternal health, immunization, non-communicable diseases, and health commodity management if adapted to local data realities.

Finally, future publications should standardize reporting. ARM studies should report data source, sample size, preprocessing steps, discretization methods, missing data handling, algorithm parameters, support and confidence thresholds, rule pruning methods, validation approach, clinical interpretation, ethical approval, and limitations. Standardization would improve reproducibility and make future systematic reviews more rigorous.

## VIII. CONCLUSION

Association rule mining has become a valuable method for healthcare knowledge discovery because it produces transparent and interpretable rules from complex clinical data. The evidence reviewed in this manuscript shows applications across clinical decision support, emergency diagnostic testing, medication-problem association mining, diagnosis-medication analysis, adverse drug reaction detection, chronic disease risk factor discovery, lifestyle behavior analysis, hospital readmission, oncology recurrence prediction, patient trajectory modelling, negative rule mining, and privacy-preserving distributed analytics. The dominant algorithm remains Apriori, but recent studies increasingly use FP-Growth, weighted and modified ARM, class association rules, causal ARM, negative ARM, and privacy-preserving approaches.

The main contribution of ARM is its ability to reveal hidden relationships that can be explained to clinicians and managers. However, ARM must be applied carefully. Healthcare rules are vulnerable to missing data, coding bias, confounding, threshold sensitivity, rule explosion, and lack of external validation. Therefore, ARM should not be treated as a standalone proof-generating method but as a decision-support and hypothesis-generating approach that requires clinical review and validation. Future research should focus on validated, explainable, privacy-preserving, and workflow-integrated ARM systems that can support safer and more efficient healthcare delivery.

## REFERENCES

1. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216. https://doi.org/10.1145/170035.170072
2. Altaf, W., Shahbaz, M., & Guergachi, A. (2017). Applications of association rule mining in health informatics: A survey. Artificial Intelligence Review, 47, 313-340. https://doi.org/10.1007/s10462-016-9483-9
3. Bayley, K. B., Belnap, T., Savitz, L., Masica, A. L., Shah, N., & Fleming, N. S. (2013). Challenges in Using Electronic Health Record Data for CER. Medical Care, 51(Supplement 8Suppl 3), S80–S86. https://doi.org/10.1097/mlr.0b013e31829b1d48
4. Budaraju, R. R., & Jammalamadaka, S. K. R. (2024). Mining negative associations from medical databases considering frequent, regular, closed and maximal patterns. Computers, 13(1), 18. https://doi.org/10.3390/computers13010018
5. Cheng, C. W., Chanani, N., Venugopalan, J., Maher, K., & Wang, M. D. (2013). icuARM-An ICU clinical decision support system using association rule mining. IEEE Journal of Translational Engineering in Health and Medicine, 1, 4400110. https://doi.org/10.1109/JTEHM.2013.2290113
6. Domadiya, N., & Rao, U. P. (2019). Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. Procedia Computer Science, 148, 303-312. https://doi.org/10.1016/j.procs.2019.01.023
7. Domadiya, N., & Rao, U. P. (2021). Privacy preserving association rule mining on distributed healthcare data: A cryptographic approach. SN Computer Science, 2, 390. https://doi.org/10.1007/s42979-021-00801-7
8. Firat Atay, F. F., Yagin, F. H., Colak, C., Elkiran, E. T., Mansuri, N., Ahmad, F., & Ardigò, L. P. (2024). A hybrid machine learning model combining association rule mining and classification algorithms to predict differentiated thyroid cancer recurrence. Frontiers in Medicine, 11, 1461372. https://doi.org/10.3389/fmed.2024.1461372
9. Gatobu, S. K., Kimwele, M. W., & Okeyo, G. (2025). Enhancing Breast Cancer Prediction Model through Association Rule Mining and Classification Techniques for underserved communities: Hybrid model for breast cancer prediction. Journal of Agriculture, Science and Technology, 24(2), 32–48. https://doi.org/10.4314/jagst.v24i2.3

10. Guillamet, G. H., Seguí, F. L., Vidal-Alaball, J., & López, B. (2023). CauRuler: Causal irredundant association rule miner for complex patient trajectory modelling. Computers in Biology and Medicine, 155, 106636. https://doi.org/10.1016/j.compbiomed.2023.106636
11. Holmes, J. H., Beinlich, J., Boland, M. R., Bowles, K. H., Chen, Y., Cook, T. S., Demiris, G., Draugelis, M., Fluharty, L., Gabriel, P. E., Grundmeier, R., Hanson, C. W., Herman, D. S., Himes, B. E., Hubbard, R. A., Kahn, C. E., Jr., Kim, D., Koppel, R., Long, Q., … Moore, J. H. (2021). Why Is the Electronic Health Record So Challenging for Research and Clinical Care? Methods of Information in Medicine, 60(01/02), 032–048. https://doi.org/10.1055/s-0041-1731784
12. Ilma, H., Notodiputro, K. A., & Sartono, B. (2023). ASSOCIATION RULES IN RANDOM FOREST FOR THE MOST INTERPRETABLE MODEL. BAREKENG: Jurnal Ilmu Matematika Dan Terapan, 17(1), 0185–0196. https://doi.org/10.30598/barekengvol17iss1pp0185-0196
13. Javeedullah, M. (2025). Integrating Health Informatics into Modern Healthcare Systems: A Comprehensive Review. Global Journal of Universal Studies, 2(1), 1–21. https://doi.org/10.70445/gjus.2.1.2025.1-21
14. Kang'ethe, S. M., & Wagacha, P. W. (2014). Extracting diagnosis patterns in electronic medical records using association rule mining. **International Journal of Computer Applications, 108**(15), 19–26. https://doi.org/10.5120/18987-0425
15. Lee, D. G., Ryu, K. S., Bashir, M., Bae, J. W., & Ryu, K. H. (2013). Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. Journal of Medical Systems, 37, 9896. https://doi.org/10.1007/s10916-012-9896-1
16. Miswan, N. H., Sulaiman, I. M., Chan, C. S., & Ng, C. G. (2021). Association rules mining for hospital readmission: A case study. Mathematics, 9(21), 2706. https://doi.org/10.3390/math9212706
17. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093. https://doi.org/10.1016/j.eswa.2012.08.028
18. Narindrarangkura, P., Alafaireet, P. E., Khan, U., & Kim, M. S. (2023). Association rule mining of real-world data: Uncovering links between race, glycemic control, lipid profiles, and suicide attempts in individuals with diabetes. Informatics in Medicine Unlocked, 42, 101345. https://doi.org/10.1016/j.imu.2023.101345
19. Odu, N. B., Prasad, R., Onime, C., & Sharma, B. K. (2022). How to implement a decision support for digital health: Insights from design science perspective for action research in tuberculosis detection. Intelligent Medicine, 2(4), 192-202. https://doi.org/10.1016/j.jjimei.2022.100136
20. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., et al. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
21. Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., et al. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160
22. Park, S. H., Jang, S. Y., Kim, H., & Lee, S. W. (2014). An association rule mining-based framework for understanding lifestyle risk behaviors. PLOS ONE, 9(2), e88859. https://doi.org/10.1371/journal.pone.0088859
23. Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., & Hadaegh, F. (2015). An application of association rule mining to extract risk pattern for type 2 diabetes using Tehran Lipid and Glucose Study database. International Journal of Endocrinology and Metabolism, 13(2), e25389. https://doi.org/10.5812/ijem.25389
24. Sarıyer, G., & Öcal Taşar, C. (2020). Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining.

Health Informatics Journal, 26(3), 1977-1993. https://doi.org/10.1177/1460458219871135

25. Wang, C. H., Nguyen, P. A., Li, Y. C., Islam, M. M., Poly, T. N., Tran, Q. V., Huang, C. W., & Yang, H. C. (2021). Improved diagnosis-medication association mining to reduce pseudo-associations. Computer Methods and Programs in Biomedicine, 207, 106181. https://doi.org/10.1016/j.cmpb.2021.106181
26. Wei, J., Dai, J., Zhao, Y., Han, P., Zhu, Y., & Huang, W. (2021). Application of association rules analysis in mining adverse drug reaction signals. Applied Sciences, 11(22), 10828. https://doi.org/10.3390/app112210828
27. Wright, A., Chen, E. S., & Maloney, F. L. (2010). An automated technique for identifying associations between medications, laboratory results and problems. Journal of Biomedical Informatics, 43(6), 891-901. https://doi.org/10.1016/j.jbi.2010.09.009
28. Wright, A., McCoy, A. B., Henkin, S., Flaherty, M., & Sittig, D. F. (2013). Validation of an association rule mining-based method to infer associations between medications and problems. Applied Clinical Informatics, 4(1), 100-109. https://doi.org/10.4338/ACI-2012-12-RA-0051
29. Yamamoto, H., Kayanuma, G., Nagashima, T., Toda, C., Nagayasu, K., & Kaneko, S. (2023). Early detection of adverse drug reaction signals by association rule mining using large-scale administrative claims data. Drug Safety, 46(4), 371-389. https://doi.org/10.1007/s40264-023-01278-4