

ORIGINAL RESEARCH

Open access

Explainable Risk Stratification Model for Prioritizing Case Management Referrals Using Prior Utilization, Chronic Disease Burden, Social Needs Documentation, Missed Appointments, and Care Gap Indicators

Thabo Nkosi^{1*}, Lerato Molefe¹, Siphso Dlamini², Ayanda Mokoena¹, Kabelo Ndlovu²

Abstract

Case management programs are intended to reduce avoidable utilization and improve coordination for patients with complex medical, social, and engagement needs. Because case management capacity is limited, health systems need prioritization tools that are both clinically sensible and transparent. Existing referral methods often rely on clinician judgment, simple utilization thresholds, or proprietary risk scores that provide limited explanation. These approaches may overlook social needs, missed appointments, and care gaps that shape patient complexity and influence whether an intervention is feasible. This article proposes an explainable machine learning model that stratifies patients by risk of future high utilization and provides patient-specific reasoning. The model is designed around prior utilization, chronic disease burden, social needs documentation, missed appointments, and care gap indicators. The conceptual architecture uses a gradient-boosted classification model with a SHAP-based post-hoc explanation layer. The model would output both a risk score and a ranked list of contributing factors for each patient considered for case management referral. Conceptually, the model would identify patients who may benefit from case management and explain why each patient was prioritized. These explanations could help case managers tailor outreach, match patients to intervention pathways, and distinguish medical complexity from social instability or disengagement. An explainable risk stratification model could turn a blind referral process into a transparent, clinically sensible prioritization workflow. Its value would depend on careful implementation, fairness monitoring, and alignment with real case manager decision-making.

Keywords Risk stratification, Social determinants of health, Explainable artificial intelligence, SHAP, Population health, Case management

*Correspondence:

Thabo Nkosi

thabo.nkosi@gmail.com

¹ Department of Healthcare Analytics and Smart Systems, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

² Department of Digital Health Engineering, Faculty of Engineering, University of the Witwatersrand, Johannesburg, South Africa

Introduction

Case management programs are often asked to support more patients than they can realistically enroll, creating a persistent mismatch between population need and available care coordination capacity. Patients with repeated emergency department visits, multiple chronic conditions, social instability, missed follow-up, or unresolved care gaps

may all appear appropriate for referral, but these signals are not always visible in a single worklist. Prior research on prenatal case management, high-cost patient prediction, and persistent high-utilization populations shows that machine learning could help organize complex patient information into a more systematic prioritization process [1-3]. For case management, the central challenge is not only identifying patients at elevated risk but also explaining why they should be prioritized over others with competing needs [4, 5].

Current referral methods may depend on clinician intuition, simple utilization thresholds, payer-based rules, or vendor-generated risk scores that do not fully explain the reasoning behind patient ranking. Threshold-based strategies can identify patients with prior high utilization, but they may miss patients whose risk is driven by social barriers, care fragmentation, or patterns of disengagement that have not yet produced repeated hospitalization. Studies of high-cost and high-need patient identification show that administrative and EHR-derived models can support population-level targeting, yet opaque outputs may limit trust when case managers must decide whom to contact first [6-8]. Concerns about bias in clinical machine learning further underscore the need for transparent referral logic rather than unexamined automation [9, 10].

Electronic health records and claims systems now contain structured signals that can support a richer case management risk model, including utilization history, chronic disease burden, medication patterns, appointment adherence, and coded or screened social needs. Risk-stratification studies using EHR, claims, medication, and social determinants data suggest that patient complexity is multidimensional and cannot be reduced to diagnosis counts alone [11-13]. ICD-10 Z-codes and social needs documentation can represent housing insecurity, food insecurity, transportation barriers, and other nonmedical constraints, although these fields may be incomplete and unevenly captured [14, 15]. Missed appointments and care gaps add another layer by indicating whether a patient is connected to care, following through with recommended services, or at risk of preventable deterioration [16].

This article proposes an explainable risk stratification model for prioritizing case management referrals using prior utilization, chronic disease burden, social needs documentation, missed appointments, and care gap indicators. The model is conceptual rather than experimental, and its purpose is to describe how an XAI system could generate a clinically interpretable ranking while supporting human judgment. SHAP-based explanations would translate model output into patient-specific drivers, allowing case managers to understand whether high risk is primarily related to utilization intensity, multimorbidity, social instability, disengagement, or unresolved care gaps. The goal is not to replace professional assessment but to create a transparent, auditable referral workflow aligned with population health management and responsible machine learning principles.

Background

Case management and high-risk patient identification

Case management targets patients whose medical complexity, service fragmentation, or social barriers create a need for structured coordination beyond routine clinical care. Traditional risk-identification tools, such as diagnosis-based risk adjustment, readmission risk screens, and utilization thresholds, can support population segmentation but may not explain which modifiable factors make a patient actionable for referral. Machine learning studies in high-cost and high-need populations suggest that broader feature sets could improve conceptual sensitivity to complex patient profiles, especially when utilization, diagnosis, and service-use patterns are combined [2, 3, 17]. However, case management prioritization also requires interpretability because a high-risk label alone does not tell a care team whether to focus on disease stabilization, social support, appointment engagement, or closure of care gaps [18, 19].

Prior utilization and chronic disease burden

Prior utilization is one of the most common signals for future resource intensity because repeated emergency department visits, hospitalizations, and specialty encounters can reflect unstable disease, fragmented care, or unmet social needs.

Chronic disease burden adds complementary information by summarizing multimorbidity and the cumulative clinical work required to manage patients over time. Studies comparing risk-stratification models and high-cost patient prediction approaches show that claims and EHR-derived utilization variables are often central to identifying patients likely to require intensive resources [4-6]. For case management, these variables should be interpreted as markers of coordination need rather than as stand-alone proof that referral will be beneficial [8, 20].

Social needs documentation and its role in complexity

Social needs documentation can reveal barriers that are not adequately captured by diagnoses or prior utilization alone. Housing insecurity, food insecurity, transportation difficulty, and related Z-code indicators may help explain why patients miss care, experience avoidable deterioration, or use emergency services when outpatient management fails. Studies of social needs, social determinants, and Z-codes show that these data can add important context for risk stratification, but they are also shaped by documentation practices and screening availability [12, 14, 15]. An explainable model should therefore treat social needs both as potential contributors to risk and as signals requiring careful interpretation because absence of documentation does not mean absence of need [21].

Missed appointments and care gap indicators

Missed appointments, appointment non-adherence, and unresolved care gaps can function as engagement and access signals in a case management model. A patient with repeated no-shows may require outreach, transportation support, or navigation assistance, while a patient with overdue monitoring or incomplete preventive care may need structured follow-up before a preventable event occurs. Machine learning studies of missed appointment prediction and medication adherence support the idea that engagement-related variables can carry operational meaning beyond simple disease severity [13, 16]. In a case management setting, these indicators would be most useful when paired with explanations that show whether risk arises from disengagement, uncontrolled disease, or both [19].

Explainable AI in population health

Explainable AI is essential in population health because prioritization tools influence access to limited services, not merely prediction labels. Case managers, clinicians, and administrators must be able to inspect why a patient is ranked highly, whether the explanation is clinically plausible, and whether the model might reproduce historical inequities in utilization or access. Work on responsible machine learning, fairness, and clinical AI implementation emphasizes that transparency, human-centered design, and ongoing audit are necessary for safe deployment [22-24]. SHAP and related explanation approaches can help translate complex model behavior into patient-specific reasoning, but the explanations must be embedded in workflows where humans can challenge, contextualize, and act on them [19, 25].

Model Development Overview

High-level prediction pipeline

The proposed system would run on a regular population health cadence, such as weekly or monthly, across an attributed patient panel. It would ingest structured EHR and claims-derived data, generate a case management referral risk score for each eligible patient, rank patients for review, and attach a SHAP-based explanation for those at the top of the worklist. This pipeline follows the logic of population-based risk stratification models that combine demographic, diagnosis, medication, utilization, and service-use information to support proactive targeting [6, 11, 26]. The output would be a decision-support worklist for case managers, not an automatic enrollment decision.

Core input features

The core feature set would include prior emergency department visits, hospitalizations, outpatient utilization patterns, chronic disease burden, social needs flags, missed appointment rate, and open care gap indicators. These domains

reflect evidence that future resource intensity can be shaped by both medical complexity and nonmedical barriers, including social determinants, medication adherence, and engagement with care [12, 13, 15]. Chronic disease burden would be represented conceptually through comorbidity indices or grouped condition categories, while utilization would be summarized through recent patterns rather than isolated events [2, 4]. The model would be designed to treat each feature group as one part of a broader patient complexity profile rather than as a single deterministic reason for referral.

Design principles

The model should be interpretable, action-focused, regularly updated, and designed to augment rather than replace case manager judgment. Its explanations should identify actionable drivers such as missed appointments, transportation barriers, medication adherence concerns, or unclosed care gaps, while also acknowledging less modifiable drivers such as age or chronic disease burden. Human-centered design research in care management emphasizes that risk tools are more likely to be useful when embedded in actual work practices and presented in language that supports decision-making [19]. Fairness and responsible AI principles further require that the model be auditable across demographic, payer, and social-risk groups before it is used to allocate scarce case management resources [9, 10, 22].

Figure 1 illustrates the end-to-end explainable risk stratification workflow for case management referral prioritization, from structured healthcare data inputs and predictor domains to model scoring, SHAP-based explanation, fairness review, case manager interpretation, and practical implementation action.

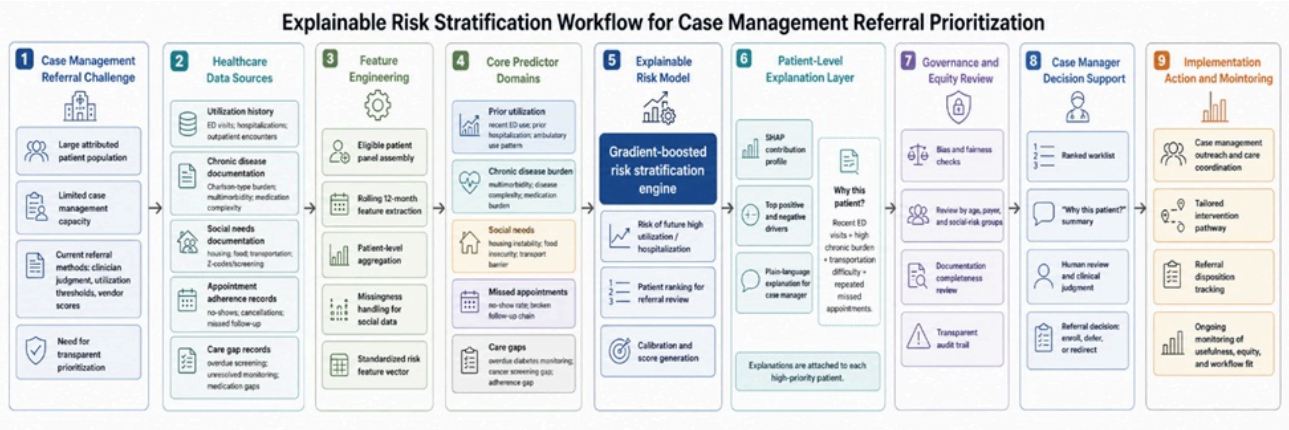


Figure 1. Explainable risk stratification workflow for prioritizing case management referrals using utilization, chronic disease burden, social needs, missed appointments, and care gaps

Data Sources and Feature Engineering

Utilization and chronic disease features

Utilization and chronic disease features would be extracted from claims feeds, encounter tables, problem lists, medication records, and EHR utilization databases. Rolling indicators would summarize recent emergency department use, hospital admissions, ambulatory visits, and specialty encounters, while comorbidity indices or grouped chronic condition counts would represent cumulative clinical burden. Prior studies comparing EHR-based and claims-based risk stratification indicate that these data sources can provide complementary perspectives on patient complexity and future resource needs [6, 11, 26]. In this conceptual model, utilization features would be engineered to support explainability by preserving clinically meaningful categories rather than collapsing all service use into an opaque aggregate score.

Social needs documentation features

Social needs features would be represented as binary flags, categorical indicators, or composite scores derived from ICD-10 Z-codes, structured screening tools, and documented barriers such as housing instability, food insecurity, and transportation difficulty. These variables would help the model distinguish patients whose risk is driven by nonmedical constraints from those whose risk is primarily driven by disease burden or recent acute care use. Studies on social needs data and Z-code use show that such indicators can add important context to utilization prediction, but they may be under-coded and unevenly documented across clinical settings [12, 14, 15]. To reduce the risk of interpreting missing documentation as true absence of need, the feature vector should include explicit missingness categories and support audit of documentation patterns across patient groups [21].

Missed appointment and care gap features

Missed appointment features would summarize patterns of no-shows, late cancellations, or incomplete follow-up across ambulatory care settings, while care gap features would indicate overdue monitoring, incomplete screening, or unresolved recommended services. These features would be designed to capture engagement and continuity problems that may precede high utilization but are not always reflected in hospitalization history. Research on appointment non-adherence and medication adherence suggests that engagement-related signals can support risk stratification when interpreted as potential barriers rather than patient blame [13, 16]. In the proposed model, these features would be paired with plain-language explanations so case managers can translate them into outreach, navigation, or care coordination actions.

Interpretable Model Architecture and Explainability Layer

Model choice and rationale

A gradient-boosted classification model would be a reasonable conceptual choice because population health risk stratification often relies on structured, mixed-type data drawn from diagnoses, utilization, medications, social documentation, and operational records. The model would be calibrated conceptually to estimate the probability of near-term high utilization or preventable hospitalization, while avoiding automatic assignment to case management without human review. Prior studies of high-cost patient prediction and persistent high-utilizer identification suggest that ensemble and tree-based approaches can represent nonlinear interactions among utilization, comorbidity, and service-use features [4, 5, 17]. For an XAI-oriented design, model choice must be paired with explanation methods and governance processes that make patient ranking inspectable and clinically contestable [24, 25].

Input feature vector construction

The patient-level input vector would aggregate structured features across clinical, utilization, social, engagement, and care gap domains. Missingness would be handled explicitly, particularly for social needs documentation, because incomplete screening or under-coding may reflect workflow variation rather than the patient's true social context. Bias research in clinical machine learning warns that EHR variables can encode historical inequities, documentation differences, and access barriers, making transparent feature construction essential [9, 10, 21]. The model should therefore preserve feature provenance so case managers and analysts can see whether a risk contribution came from observed utilization, documented need, missed care, or missing data.

SHAP explanation generation and integration

For each high-priority patient, the explainability layer would generate a SHAP-based summary showing the leading factors that increased or decreased the patient's risk score. A case manager-facing view could pair a compact visual explanation with a plain-language statement such as recent emergency visits, uncontrolled chronic disease burden, documented transportation difficulty, and missed appointments contributed to prioritization. SHAP has been used to connect local explanations with broader model understanding, making it conceptually appropriate for translating tree-based risk models into patient-specific reasoning [25]. To be useful in case management, the explanation should be

embedded directly in the referral worklist and written in operational language that supports outreach planning rather than technical model interpretation alone [19].

Integrating clinical, utilization, social, and engagement signals

Table 1 summarizes the major predictor domains in the proposed model and shows how each domain contributes to patient-level explanation and case management action.

Table 1. Predictor domains, operational meaning, and explanation outputs in the explainable risk stratification model for case management referral prioritization

Predictor domain	Representative variables in the manuscript	Why the domain matters for referral prioritization	What a high-risk signal would indicate conceptually	Example explanation message a case manager could receive	Practical case management implication
Prior utilization	Prior emergency department visits, prior hospitalizations, outpatient utilization pattern	Historical utilization is a strong indicator of instability, fragmented care, or unresolved health needs requiring coordination	Recurrent or recent acute care use may suggest elevated near-term risk and unmet care management need	"Recent emergency and inpatient utilization contributed strongly to this patient's priority ranking."	Prioritize review for transitional care, care coordination, or intensified follow-up after recent acute episodes
Chronic disease burden	Multimorbidity, chronic disease burden score, medication complexity	High disease burden reflects clinical complexity, need for coordination, and risk of resource-intensive care	Multiple chronic conditions or treatment complexity may increase the need for proactive management	"High chronic disease burden and complex ongoing management needs increased this patient's referral priority."	Consider nurse case management, medication reconciliation, and longitudinal chronic disease support
Social needs documentation	Housing instability, food insecurity, transportation barriers, Z-code or screening-based social flags	Social needs may drive missed care, preventable deterioration, and barriers to successful self-management	Nonmedical barriers may amplify clinical risk and reduce the chance that usual care alone will be effective	"Documented transportation and housing-related barriers were important contributors to this patient's risk profile."	Route toward community resource linkage, social work support, or navigation assistance

Missed appointments	No-show rate, missed follow-up visits, repeated cancellation pattern	Missed appointments can signal poor engagement, access difficulty, or care disruption before avoidable deterioration occurs	Repeated nonattendance may indicate a need for outreach, engagement support, or barrier assessment	"Frequent missed appointments contributed to this patient's prioritization for case management review."	Trigger outreach, reminder support, transportation assessment, or community health worker engagement
Care gap indicators	Overdue cancer screening, unresolved diabetes monitoring, medication adherence gaps, incomplete preventive care	Care gaps identify patients who are not receiving recommended follow-up and may be at risk of worsening outcomes	Unresolved preventive or chronic care gaps may indicate care fragmentation or ineffective follow-through	"Multiple open care gaps increased this patient's priority for proactive care coordination."	Support gap closure planning, preventive outreach, and coordination with primary care workflows
Combined medical and social interaction	High utilization plus social needs, multimorbidity plus nonadherence, care gaps plus missed appointments	Risk may arise from interactions across domains rather than a single variable	The patient's complexity may reflect overlapping clinical instability and social or behavioral barriers	"This patient ranked highly because recent acute care use co-occurred with social barriers and repeated missed care."	Assign to a more intensive or multidisciplinary case management pathway
Missingness-aware interpretation	Explicit missingness categories for social screening or incomplete documentation	Lack of documentation can distort risk estimation and obscure true patient need	Missing social data may reflect workflow limitations, not absence of barriers	"Social need fields were incompletely documented; interpretation should be combined with clinical review."	Encourage case manager verification, supplemental screening, and cautious use of incomplete fields
SHAP explanation layer	Ranked feature contributions attached to each patient	Explanations convert model output into patient-specific reasons that support action	The model does not only rank patients; it shows why each patient was ranked	"Top drivers of risk included prior ED use, chronic disease burden, transportation difficulty, and missed visits."	Improves trust, supports triage, and helps align referral choice with the underlying reason for risk

Combining medical and social risk

The proposed model would integrate medical and social risk by allowing prior utilization, chronic disease burden, and documented social needs to contribute jointly to referral prioritization. For example, repeated acute care use combined with housing instability or transportation barriers could indicate a patient whose risk is not simply clinical but also shaped

by difficulty accessing continuous outpatient care. Studies of behavioral health factors, social determinants, and high-cost utilization suggest that cost and utilization risk often emerge from interacting clinical and social conditions rather than from a single diagnosis or service-use variable [12, 27, 28]. In this setting, explainability would help case managers distinguish patients whose risk reflects medical escalation from those whose risk reflects unmet social needs that may be amenable to navigation or community-resource linkage.

Distinguishing high-risk by engagement vs. Disease

Two patients may receive similar priority scores for very different reasons, making explanation essential for case management workflow. One patient's risk may be driven mainly by advanced multimorbidity, medication complexity, and recent hospitalization, while another patient's risk may be driven by missed appointments, unclosed care gaps, and social documentation suggesting difficulty remaining connected to care. Prior work on medication adherence, missed appointments, and high-risk patient prediction supports the idea that engagement-related signals should be interpreted differently from disease-severity signals [13, 16, 20]. A patient-specific explanation would allow the case manager to decide whether the appropriate response is clinical coordination, outreach support, transportation assistance, medication reconciliation, or escalation to a multidisciplinary care team.

Weighting the recency of utilization

The model would be expected to treat recent utilization as more informative than distant utilization when recent events suggest active decompensation, instability, or failed outpatient management. A conceptual feature-engineering strategy could therefore summarize emergency department visits, hospitalizations, and ambulatory encounters using recency-weighted windows rather than treating all historical events equally. Research on persistent high utilizers and high-cost patient prediction shows that utilization histories can help identify patients who may remain resource-intensive, but the operational meaning of those histories depends on whether use is recent, sustained, episodic, or declining [3, 5, 8]. Recency weighting would make the risk score more responsive to current care coordination needs while preserving explanations that show which utilization patterns influenced ranking.

Adjusting for age and payer type

Age, payer type, and access-related variables can improve risk stratification, but they also require careful fairness review because they may encode structural differences in coverage, access, and historical care use. A fairness-aware implementation would examine whether the model systematically under-prioritizes patients who have fewer documented encounters because of access barriers rather than lower need. Literature on algorithmic fairness, racial bias in population health algorithms, and disparities in clinical AI warns that risk models can reproduce inequities when utilization is treated as a neutral proxy for need [10, 21, 22]. Explanations should therefore be audited by demographic and payer groups to determine whether the model is ranking patients because of clinically meaningful risk or because of biased measurement patterns.

Explainability for Case Manager Workflow

Case manager-facing risk report

The case manager-facing output would be a daily or weekly referral worklist showing top-ranked patients, a concise risk category, and a short “why this patient?” explanation. Instead of presenting a score alone, the report would identify the major contributors, such as recent emergency department use, multiple chronic conditions, documented transportation difficulty, missed visits, or overdue monitoring. Human-centered design research in care management emphasizes that machine learning outputs must fit the user's workflow, vocabulary, and decision responsibilities to become practically useful [19]. SHAP explanations could support this design by translating model behavior into local, patient-level drivers that case managers can review before making referral decisions [25].

Actionable subgroups and tailored interventions

The model could support actionable subgrouping by distinguishing patients whose high-risk profile is dominated by poorly controlled disease, social instability, care disengagement, or unresolved care gaps. Such subgrouping would not be a claim that the model has discovered fixed patient types; rather, it would be a workflow aid that helps case managers choose an initial intervention pathway. Studies of high-cost users, clustering, and population risk segmentation suggest that patients with similar utilization levels may still differ in the underlying drivers of complexity [8, 17, 28]. Explanations would be central because they would show whether the subgroup label is supported by documented clinical burden, service-use patterns, social barriers, or engagement indicators.

Counterfactual reasoning for engagement

Counterfactual reasoning could help case managers understand how modifiable engagement factors might influence a patient's risk profile without presenting the estimate as a guaranteed outcome. For example, the system could indicate that missed appointments or unresolved monitoring gaps are important contributors and that improved follow-up would be expected to lower concern, while avoiding precise numerical claims. Research on missed appointment prediction and care management implementation supports the value of translating engagement signals into practical outreach actions rather than treating non-attendance as patient failure [16, 19]. In this model, counterfactual explanations would be framed as decision-support prompts for navigation, reminder systems, community health worker outreach, or barrier assessment.

Audit and feedback loop

Case managers should be able to record whether an explanation was clinically sensible, whether the patient was appropriate for referral, whether contact was successful, and whether an alternative intervention was more suitable. This feedback would support model governance by identifying recurring false prioritization patterns, documentation gaps, or explanation wording that does not match real care management reasoning. Responsible machine learning frameworks emphasize that clinical AI should be monitored after deployment, reviewed by users, and revised when model behavior conflicts with safety, equity, or workflow expectations [18, 23, 24]. Feedback should refine the model and explanation interface over time while maintaining human accountability for referral decisions.

Operational Deployment and Referral Integration

Integration with care management platforms

The risk score and explanation should be delivered within the case management platform or EHR work queue where referral decisions are already made. A useful implementation would allow one-click referral, documentation of the reason for referral, and linkage to relevant patient context such as recent utilization, active care gaps, and documented social needs. Implementation research in clinical AI and care management shows that risk tools are more likely to matter when they are embedded into existing work rather than placed in separate dashboards that require additional effort [19, 23]. The model should therefore be treated as an operational decision-support layer that organizes information for review, not as an external analytic product disconnected from daily work.

Creating a closed-loop referral system

A closed-loop referral system would track whether high-priority patients were reviewed, referred, contacted, enrolled, declined, or redirected to another service. These outcomes could inform future model monitoring by showing whether the worklist is identifying patients who are both high-risk and realistically reachable through case management. Population health management studies emphasize that risk stratification must be connected to action, because prediction without a feasible intervention pathway may not improve care coordination [20, 26]. The system should therefore learn from referral disposition and case manager feedback while avoiding automatic reinforcement of biased outreach patterns or exclusion of patients who are difficult to contact [9, 10].

Table 2 presents a practical framework for fairness, transparency, implementation, and ongoing monitoring of the explainable case management referral model.

Table 2. Fairness, transparency, implementation, and monitoring framework for operationalizing the explainable case management referral model

Implementation domain	Operational purpose	Key question for the health system	Recommended practical approach	Expected value for deployment
Transparency of patient ranking	Make referral prioritization understandable to end users	Why was this patient ranked highly for case management review?	Display the risk score with a concise SHAP-based “why this patient?” summary in the same workflow screen	Improves interpretability and reduces blind reliance on a numeric score
Human-centered review	Preserve clinician and case manager judgment	How should model output be used without replacing professional assessment?	Require human review before referral action and allow users to enroll, defer, or redirect patients	Keeps the model as decision support rather than automated determination
Fairness assessment	Detect inequitable prioritization patterns	Does the model under-prioritize or over-prioritize certain demographic, payer, or social-risk groups?	Review model behavior and explanation patterns across age, payer type, and social-risk strata before and during deployment	Supports equitable allocation of limited case management resources
Data completeness review	Address documentation bias	Are social needs and engagement variables captured consistently enough to support fair interpretation?	Audit missingness patterns, under-coding of social needs, and variation in screening uptake	Reduces the risk of treating missing data as absence of need
Workflow integration	Ensure the tool fits real care management operations	Can users act on model output without leaving their existing workflow?	Embed the ranked worklist and explanation in the EHR or case management platform with one-click referral functionality	Increases practical usability and reduces workflow burden
Actionability of outputs	Link prediction to intervention	Does the model output support a meaningful next step for the case manager?	Pair explanations with likely intervention categories such as disease stabilization, social support, engagement outreach, or care gap closure	Moves the system from prediction to operational action
Referral tracking	Close the operational loop	What happened after a patient was prioritized?	Record whether the patient was reviewed, referred, contacted, enrolled, declined, or redirected	Enables assessment of real-world usefulness and operational follow-through
Monitoring of explanation usefulness	Determine whether explanations actually help users	Do case managers find the explanations	Collect structured user feedback on explanation	Improves the explanation interface and supports adoption

		clear, plausible, and actionable?	clarity, trustworthiness, and decision relevance	
Governance and oversight	Provide accountability for a high-impact workflow	Who is responsible for reviewing model behavior and resolving concerns?	Establish oversight involving care management leadership, informatics, and quality/governance teams	Strengthens accountability and responsible implementation
Model updating and monitoring	Maintain relevance over time	Does the model remain aligned with changing workflows, documentation, and patient populations?	Periodically review input patterns, prioritization behavior, fairness indicators, and user feedback	Supports sustainable, safe, and context-aware deployment
Failure mode management	Anticipate misuse or misinterpretation	What if the model ranks patients for the wrong reasons or explanations are misleading?	Flag unusual explanation patterns, review edge cases, and maintain manual override capacity	Reduces operational harm and prevents overdependence on the tool
Implementation success evaluation	Assess whether the model improves referral practice	Does the system make case management referral more transparent, targeted, and efficient?	Evaluate referral appropriateness, engagement success, workload impact, and equity of prioritization	Demonstrates whether the model adds practical value beyond existing referral processes

Evaluation Strategy

Predictive performance and fairness

The model should be evaluated using standard discrimination, calibration, and precision-recall concepts, but the emphasis should be on whether risk estimates are reliable enough to support referral prioritization rather than on isolated performance numbers. Fairness evaluation should examine calibration and prioritization patterns across demographic, payer, language, and social-risk groups to determine whether the model is systematically over- or under-identifying particular populations. Prior studies on fairness, racial bias in population health algorithms, and clinical machine learning governance show why performance assessment must include equity review, not only aggregate predictive accuracy [10, 22, 24]. The evaluation should also test whether explanations differ systematically across groups in ways that might reveal biased feature use or documentation inequity.

Explanation quality and usefulness

Explanation quality should be evaluated by asking case managers whether the model's reasons are clear, clinically plausible, actionable, and consistent with their understanding of the patient. The assessment should include whether explanations help users distinguish disease complexity, social instability, missed appointments, and care gaps as different pathways to risk. Research on explainable tree-based models and user-centered machine learning design supports evaluating explanations in the context of actual decision-making rather than assuming that a technical explanation is automatically useful [19, 25]. The goal would be to determine whether explanations improve trust calibration, referral appropriateness, and intervention selection without encouraging uncritical acceptance of the model.

Impact on outcomes and resource allocation

The model's operational value should be evaluated through pragmatic implementation research comparing model-guided referral prioritization with usual referral processes, while avoiding premature claims of benefit before real-world testing. Relevant outcomes could include avoidable acute care use, timely case management review, successful engagement, closure of care gaps, and case manager workload, interpreted as domains for future evaluation rather than reported results. Studies of high-cost patient identification, care management prediction models, and population health risk stratification show that predictive tools must be linked to resource allocation decisions to determine whether they actually improve care delivery [1, 2, 20]. Evaluation should also examine whether limited case management capacity is distributed more transparently and equitably when explanations accompany patient ranking.

Limitations

Data quality and completeness of social data

A major limitation is that social needs documentation may be incomplete, inconsistently coded, or unevenly collected across clinics and patient groups. ICD-10 Z-codes and structured social screening fields can indicate important barriers, but under-coding may cause the model to miss patients whose social needs are real but undocumented. Studies of Z-code use and social needs data emphasize that these variables should be interpreted as documentation signals as well as patient complexity signals [12, 14, 15]. The model would therefore require missingness indicators, documentation audits, and careful explanation wording so users do not mistake absent social data for absence of social risk.

Risk of reinforcing systemic disparities

The model could reinforce systemic disparities if historical utilization, access, payer status, or documentation patterns are used without fairness safeguards. Patients with limited access to care may appear lower risk because they have fewer recorded visits, while patients from groups subject to biased documentation may be ranked for reasons that reflect system behavior rather than true need. Prior work on bias, fairness, and responsible clinical machine learning demonstrates that algorithmic tools can reproduce inequitable patterns unless they are audited before and after deployment [9, 10, 22]. For this reason, the model should be implemented with ongoing fairness monitoring, human review, and governance procedures that allow correction when prioritization patterns appear unjustified.

Conclusion

An explainable risk stratification model for case management referral prioritization could help health systems organize complex patient information into a transparent, clinically meaningful worklist. By integrating prior utilization, chronic disease burden, social needs documentation, missed appointments, and care gap indicators, the model could identify patients whose risk reflects multiple overlapping forms of complexity. Its purpose would be to support prioritization under limited case management capacity, not to replace professional judgment or automatically determine enrollment.

The key strength of the proposed approach is its combination of multidimensional risk modeling with patient-specific explanation. A SHAP-based explanation layer could show why a patient ranks highly and help case managers decide whether the likely intervention should focus on disease stabilization, social support, appointment engagement, medication adherence, or care gap closure. Explicit fairness review would further support trust by making it possible to examine whether the model's recommendations are consistent, clinically sensible, and equitable across patient groups.

Important challenges remain before such a model could be responsibly implemented. Social needs data may be incomplete, care gap indicators may vary by documentation workflow, and utilization-based features may reflect access inequities as much as underlying clinical need. Operational integration may also be difficult if case managers experience the worklist as another administrative burden rather than a tool that clarifies and accelerates prioritization.

Future work should focus on pragmatic implementation trials and co-design with case managers, nurses, social workers, primary care teams, and population health leaders. The central question should not be whether an explainable model can produce a ranked list, but whether that list improves the fairness, transparency, and usefulness of referral decisions. A successful system would balance predictive accuracy with human-centered explanation, accountable governance, and practical alignment with the realities of case management.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jul 2024

Revised: 17 Aug 2024

Accepted: 24 Sep 2024

Published online: 25 February 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Pan I, Nolan LB, Brown RR, Khan R, van der Boor P, Harris DG, et al. Machine learning for social services: a study of prenatal case management in Illinois. *Am J Public Health*. 2017;107(6):938-44.
- Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high-cost high-need patient expenditures in health care. *Biomed Eng Online*. 2018;17(Suppl 1):131.
- Ng SH, Rahman N, Ang IY, Sridharan S, Ramachandran S, Wang DD, et al. Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. *BMJ Open*. 2020;10(1):e031622.
- Hu L, Li L, Ji J, Sanderson M. Identifying and understanding determinants of high healthcare costs for breast cancer: a quantile regression machine learning approach. *BMC Health Serv Res*. 2020;20(1):1066.

Howson SN, McShea MJ, Ramachandran R, Burkom HS, Chang HY, Weiner JP, et al. Improving the prediction of persistent high health care utilizers: retrospective analysis using ensemble methodology. *JMIR Med Inform.* 2022;10(3):e33212.

Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care.* 2017;55(8):789-96.

Wang Y, Iyengar V, Hu J, Kho D, Falconer E, Docherty JP, et al. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. *Front Psychiatry.* 2017;8:114.

Ramachandran R, McShea MJ, Howson SN, Burkom HS, Chang HY, Weiner JP, et al. Assessing the value of unsupervised clustering in predicting persistent high health care utilizers: retrospective analysis of insurance claims data. *JMIR Med Inform.* 2021;9(11):e31442.

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* 2018;178(11):1544-7.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-53.

Chang HY, Richards TM, Shermock KM, Dalpoas SE, Kan HJ, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. *Med Care.* 2017;55(12):1052-60.

Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. *Popul Health Manag.* 2021;24(2):222-30.

Kharrazi H, Ma X, Chang HY, Richards TM, Jung C. Comparing the predictive effects of patient medication adherence indices in electronic health record and claims-based risk stratification models. *Popul Health Manag.* 2021;24(5):601-9.

Inouye CF, Kaminski MA. Z codes for social determinants of health: a better predictor of risk for value-based care. *Popul Health Manag.* 2022;25(6):831-3.

Liss DT, Kang RH, Cherupally M, Cooper AJ, Barreto-Parra PN, Aikman C, et al. Association between ICD-10 codes for social needs and subsequent emergency and inpatient use. *Med Care.* 2024;62(1):60-6.

Yang Y, Madanian S, Parry D. Enhancing health equity by predicting missed appointments in health care: machine learning study. *JMIR Med Inform.* 2024;12(1):e48273.

Langenberger B, Schulte T, Groene O. The application of machine learning to predict high-cost patients: a performance comparison of different models using healthcare claims data. *PLoS One.* 2023;18(1):e0279540.

Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *AMIA Summits Transl Sci Proc.* 2020;2020:191-200.

Seneviratne MG, Li RC, Schreier M, Lopez-Martinez D, Patel BS, Yakubovich A, et al. User-centred design for machine learning in health care: a case study from care management. *BMJ Health Care Inform.* 2022;29(1):e100656.

de Ruijter UW, Kaplan ZR, Eijkenaar F, Maas CC, van der Heide A, Bax WA, et al. Identifying persistent high-cost patients in the hospital for care management: development and validation of prediction models. *BMC Health Serv Res.* 2024;24(1):1469.

Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med.* 2020;26(1):16-7.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* 2018;169(12):866-72.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019;25(9):1337-40.

Scott M Lundberg, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2(1):56-67.

Pandya CJ, Chang HY, Kharrazi H. Electronic health record-based risk stratification: a potential key ingredient to achieving value-based care. *Popul Health Manag.* 2021;24(6):654-6.

Sterling S, Chi F, Weisner C, Grant R, Pruzansky A, Bui S, et al. Association of behavioral health factors and social determinants of health with high and persistently high healthcare costs. *Prev Med Rep.* 2018;11:154-9.

Nghiem N, Atkinson J, Nguyen BP, Tran-Duy A, Wilson N. Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets. *Health Econ Rev.* 2023;13(1):9.