

Leveraging AI-Driven Predictive Modeling for Trauma-Informed Integrated Care: A Multicenter Implementation Study in Underserved Urban Settings

John Smart Mgbakor

Abstract: Trauma exposure is common, unevenly distributed, and strongly associated with chronic disease, mental illness, and avoidable acute care, yet integrated care programs in underserved urban settings struggle to identify patients who need outreach before a crisis occurs. Artificial intelligence (AI) can help prioritize outreach, but predictive models trained on health-system data can reproduce, and even amplify, the access barriers that define underserved populations. This paper presents a multicenter implementation framework for AI-driven predictive modeling in trauma-informed integrated care and a technical validation of its predictive and equity-audit components. The framework aligns model design with trauma-informed principles and a hybrid stepped-wedge evaluation. The technical validation uses a synthetic cohort of 16,000 patients across six simulated urban sites, built to mimic access-dependent label bias; it demonstrates the method and is not an empirical finding about real patients. A model trained on recorded crisis contacts reached 30.4% of true-need patients with limited access to care at a 10% outreach capacity, compared with 47.1% of patients with adequate access (gap 16.7 percentage points; 95% CI 13.1 to 20.3), even though its AUC was 0.83. Training on clinically adjudicated need reduced the access gap to 2.3 points (95% CI -1.8 to 6.1), and need-proportional allocation redistributed outreach across deprivation strata. In leave-one-site-out validation, AUC ranged from 0.81 to 0.84 but calibration intercepts ranged from -0.22 to $+0.49$, so local recalibration is essential. Safe use of AI in trauma-informed care depends on label choice, equity auditing, recalibration, and governance more than on algorithm choice, and a real multicenter study must now measure the implementation outcomes specified here.

Introduction

Exposure to trauma, including violence, abuse, neglect, serious accidents, forced displacement, and chronic adversity, is a common feature of human experience and a major determinant of later health. The Substance Abuse and Mental Health Services Administration (SAMHSA) defines trauma as an event, series of events, or set of circumstances experienced as physically or emotionally harmful or threatening, with lasting adverse effects on functioning and well-being [1]. The Adverse Childhood Experiences (ACE) Study showed a graded relationship between the number of childhood adversities and adult risk behaviors and disease [2], and a later meta-analysis confirmed increased risks, particularly at four or more adversities [3]. Exposure and access to treatment are not distributed evenly: they differ across social groups and are shaped by neighborhood conditions, discrimination, and the availability of services [4], [5].

2041 Washington Avenue, Second Floor, Bronx, NY 10457, United States

Email: jsmart@wholewardhealth.com

Keywords: Trauma-Informed Care, Integrated Care, Predictive Modeling, Artificial Intelligence, Health Equity, Algorithmic Fairness, Implementation Science, Underserved Urban Populations, Digital Health

Integrated care aims to respond to this complexity by coordinating physical health, behavioral health, and social support around the whole person. Figure 1 illustrates the logic of such a model: a person with complex health needs sits at the centre, supported by family and caregivers, a coordinated healthcare system, and a collaborative community, with cognitive health, social and behavioral determinants, life course, system integration, and health policy shaping resilience. In underserved urban neighborhoods every layer is under strain, so programs must decide, with limited staff time, which patients to contact first and do so before a crisis.

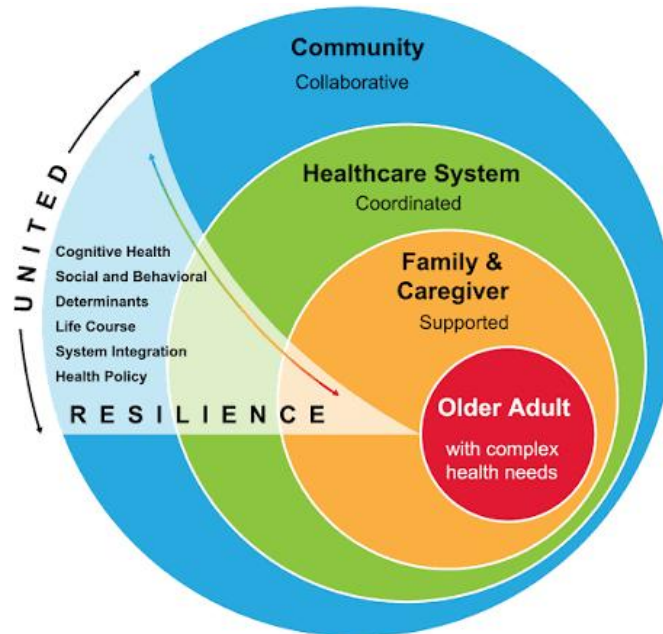


Figure 1: Conceptual model of person-centred integrated care, shown for an older adult with complex health needs at the centre, surrounded by supported family and caregivers, a coordinated healthcare system, and a collaborative community, with cognitive health, social and behavioral determinants, life course, system integration, and health policy contributing to united resilience.

Artificial intelligence (AI) offers a means of prioritizing outreach. Predictive models can combine screening results, social needs, and health-system data to estimate the risk that a patient will experience a trauma-related crisis in the near future, and the resulting ranked list can direct care-team attention. In this paper, a trauma trigger is understood as a person-specific combination of exposures, reminders, and stressors that precipitates acute distress or disengagement from care, and the prediction target is operationalized as a trauma-related crisis within 90 days. The aim is to enable earlier, more personal, and more supportive contact. It is not to monitor, label, or sanction patients.

The central risk is that prediction in underserved settings may widen the disparities it is meant to narrow. A widely cited analysis of a commercial risk algorithm found that, because it used health care costs as a proxy for health needs, it underestimated the needs of Black patients; correcting the proxy would have raised the share of Black patients flagged for additional help from 17.7% to 46.5% [6]. In underserved urban settings, people with unmet need frequently generate fewer recorded encounters because they face barriers to care. A model trained on recorded encounters will then learn that low utilization means low need. This risk is especially serious for trauma-informed care, in which a poorly timed or intrusive contact can itself be re-traumatizing.

This study builds on prior work on national trauma-informed frameworks and underserved settings by the author, and asks three questions. First, how should an AI-supported integrated care program be designed so that its operation is consistent with the principles of trauma-informed care? Second, how do the choice of training label and the allocation of limited outreach capacity affect equity of reach across patients with different levels of access to care? Third, what implementation outcomes and governance arrangements are needed to evaluate such a program across multiple sites? The contributions are: (i) a multicenter implementation framework that translates trauma-informed principles into model and workflow requirements; (ii) a reproducible technical validation, using a

synthetic multicenter cohort, of the predictive and equity-audit components; and (iii) a pre-specified implementation evaluation plan for a stepped-wedge study. Sections II to VIII cover the literature, related work, methodology, validation setup, results, discussion, and conclusion.

Literature Review

A. Trauma, Adverse Experiences, and Trauma-Informed Care

Trauma-informed care is an organizational approach, not a single treatment. SAMHSA describes six principles: safety; trustworthiness and transparency; peer support; collaboration and mutuality; empowerment, voice, and choice; and attention to cultural, historical, and gender issues [1]. Evidence for the effectiveness of trauma-informed organizational interventions is still limited: a systematic review of interventions that included staff training found a small and methodologically weak evidence base, with few rigorous evaluations of patient outcomes [7].

B. Integrated Care, Social Determinants, and Screening

Integrated care frameworks emphasize coordination across clinical, organizational, and community levels [8]. Collaborative care, in which primary care is supported by care managers and psychiatric consultation, improves depression and anxiety outcomes [9] and has improved depression and disease control in patients with chronic illness [10]. Social determinants, described as the causes of the causes, shape both exposure to trauma and the ability to benefit from care [5]. Brief validated instruments make systematic identification feasible in routine practice: the PHQ-9 for depression severity [11], the PC-PTSD-5 for probable post-traumatic stress disorder in primary care [12], and standardized tools for health-related social needs [13]. These instruments provide the structured inputs on which a predictive model for trauma-informed care can be built.

C. Predictive Modeling in Behavioral and Integrated Care

Prediction models in behavioral health have concentrated on suicide and treatment response. A model using electronic health record data identified a small high-risk fraction of outpatient visits that accounted for a large share of subsequent suicide attempts and deaths [14], but a systematic review found that such models, despite adequate discrimination, had modest positive predictive value when the outcome was rare, which limits their clinical usefulness [15]. More generally, a model's performance in one setting cannot be assumed in another: an externally validated sepsis model performed far below its developer's reports [16], and reviews of clinical AI repeatedly stress that implementation, workflow integration, and evaluation of outcomes determine clinical impact [17].

The experience of imaging illustrates how clinical AI matures when it is embedded at defined workflow touchpoints. As Figure 2 shows, algorithms now support the whole imaging pathway: machine learning decision algorithms at image prescription, automated protocoling at the study protocol stage, deep learning reconstruction during acquisition, AI feature detection, segmentation, and classification at interpretation, and AI-built structured reports at dictation. The lesson for trauma-informed care is that a risk score on its own is not an intervention. Value arises when predictions are connected to specific steps, such as intake, risk stratification, outreach protocol, the clinical encounter, and documentation, each with a responsible human role.

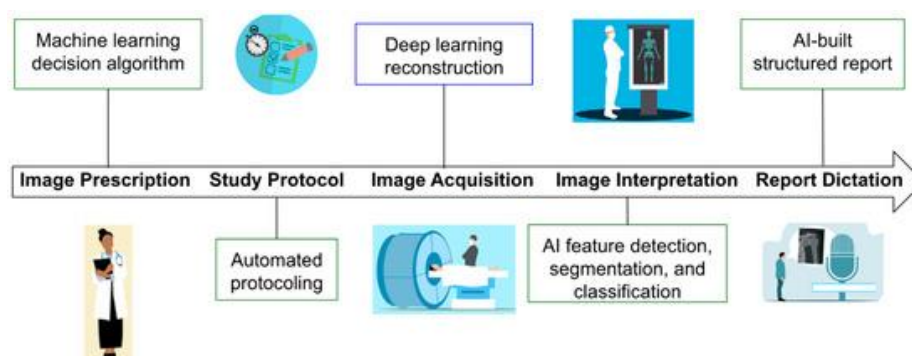


Figure 2: AI across the imaging care pathway, showing a machine learning decision algorithm at image prescription, automated protocoling at study protocol, deep learning reconstruction at image acquisition, AI feature detection, segmentation, and classification at image interpretation, and an AI-built structured report at report dictation.

D. Algorithmic Fairness and Health Equity

Bias can enter a clinical algorithm through the choice of label, the representativeness of the data, missing or mismeasured predictors, and the way outputs are used [18], [19]. The cost-based algorithm described above is the best-known example of label bias [6]. Underdiagnosis in chest radiograph classifiers was found to be higher in under-served subgroups, including patients insured through Medicaid and younger, female, Black, and Hispanic patients [20], showing that performance gaps arise in widely used systems built on large datasets. Fairness criteria such as equality of opportunity, which requires equal true positive rates across groups [21], translate naturally into a clinical question: do patients with equal need have an equal chance of being identified? Guidance for health equity stresses that fairness must be addressed throughout the model life cycle [22].

Related work

Three strands of related work are relevant. The first concerns responsible deployment. Wiens and colleagues proposed a roadmap for responsible machine learning in health care that links problem selection, development, validation, deployment, and monitoring, and emphasizes that models must be built with clinicians and tested for their effects on care [23]. The WHO guidance on the ethics and governance of AI for health adds principles of autonomy, safety, transparency, accountability, inclusiveness, and sustainability [24]. The second strand is implementation science. The Consolidated Framework for Implementation Research (CFIR) organizes the determinants of implementation across intervention, setting, people, and process [25]; RE-AIM evaluates reach, effectiveness, adoption, implementation, and maintenance [26]. Hybrid effectiveness–implementation designs allow clinical and implementation outcomes to be studied together [27], and stepped-wedge cluster designs allow every site to receive the intervention while preserving a comparison period [28].

The third strand concerns multicenter learning. Federated learning trains models across institutions without moving patient-level data [29], and has been proposed as a route to larger, more diverse training data in health care while respecting privacy [30].

Beyond prediction, AI is increasingly used to personalize care. Figure 3 summarizes four such functions: medication optimization, in which AI recommends precise medication adjustments; treatment prediction, which estimates patient responses to treatments; device therapy stratification, which identifies candidates for device therapies; and multimorbidity management, which handles interactions and adjustments among comorbid conditions. For trauma-affected patients, who frequently live with several physical and behavioral conditions, these functions represent a downstream customization layer for whole-person care. They also inherit every fairness and safety question raised by risk prediction, because a personalized recommendation is only as equitable as the data and labels behind it.

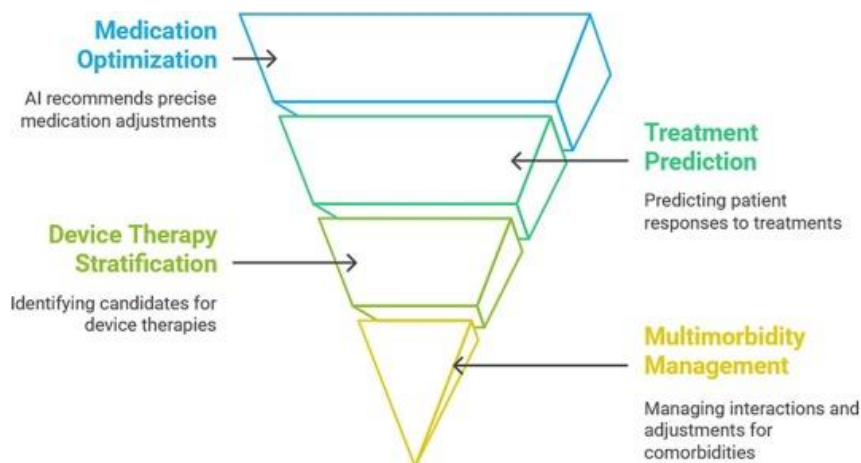


Figure 3: AI functions supporting personalized treatment: medication optimization (recommending precise medication adjustments), treatment prediction (predicting patient responses to treatments), device therapy stratification (identifying candidates for device therapies), and multimorbidity management (managing interactions and adjustments for comorbidities).

Four gaps follow. First, predictive modeling studies seldom specify how trauma-informed principles constrain model design and use. Second, equity audits often examine demographic categories after

the fact and rarely model the access mechanism that produces label bias. Third, implementation studies of clinical AI are scarce, and those that exist seldom report implementation outcomes. Fourth, evaluation tends to stop at discrimination, although reach, calibration, and workload determine whether a model helps. This paper addresses these gaps with an integrated framework and an equity-focused technical validation.

Methodology

A. Study Design and Framework

The study has two components that are distinguished throughout the paper. The first is a multicenter implementation design, specified here as a protocol for six urban sites such as federally qualified health centers and safety-net clinics. It uses a hybrid effectiveness–implementation design [27] with a stepped-wedge rollout in which sites cross from usual care to AI-supported outreach in randomized order [28]. CFIR guides the assessment of implementation context [25], and RE-AIM guides outcome evaluation [26]. This component has not yet been conducted, and no real-world implementation results are reported. The second component is a technical validation of the predictive and equity-audit pipeline on a synthetic six-site cohort (Section V), which is executed and reported. The synthetic data allow a controlled demonstration of how label choice and allocation affect equity, not a prediction of real clinics.

B. Trauma-Informed Design Requirements

The six principles of a trauma-informed approach [1] were translated into concrete requirements for the AI-supported workflow (Table 1). Each requirement is paired with a measurable check so that adherence can be audited during implementation, in keeping with the recommendation that organizational interventions be evaluated, not assumed to work [7]. Cultural humility, the commitment to self-reflection and to learning from patients and communities, was adopted as a guiding stance for outreach teams [31].

TABLE 1. Translation of trauma-informed principles into AI design requirements

Principle	Design requirement for AI-supported care	Audit check
Safety	Outreach scripts reviewed for re-traumatization risk; contact by preferred channel and time; no unsolicited disclosure of risk status	Patient-reported safety; complaint log
Trustworthiness and transparency	Plain-language notice of how data are used; patients may ask why they were contacted and may opt out	Opt-out rate; notice comprehension
Peer support	Peer workers part of the outreach team; flagged patients offered peer contact first where acceptable	Share of outreach led by peers
Collaboration and mutuality	Risk list is a prompt for conversation; clinician and patient jointly decide on actions	Documented shared decisions
Empowerment, voice, and choice	Patients can decline contact, correct data, and choose services; community advisory board co-governs	Decline rate; board decisions logged
Cultural, historical, and gender issues	Language access, culturally matched staff, and subgroup performance audits	Subgroup reach and calibration reports

C. Implementation Architecture

Figure 4 shows the proposed architecture. Data from electronic health records, validated screening instruments, social-needs screening, and utilization feed a governed feature pipeline that harmonizes definitions across sites, handles missing screening data, and applies privacy controls. The predictive model is trained on a need-based label, calibrated, and recalibrated for each site. An equity audit examines reach and calibration by access stratum and allocates outreach capacity. The care team receives a ranked outreach list with plain-language drivers and contacts patients using a trauma-informed protocol. A monitoring and governance layer tracks drift and subgroup performance, records incidents, involves a community advisory board, and triggers recalibration. Where data-sharing agreements prevent pooling, the model can be trained by federated learning [29], [30].

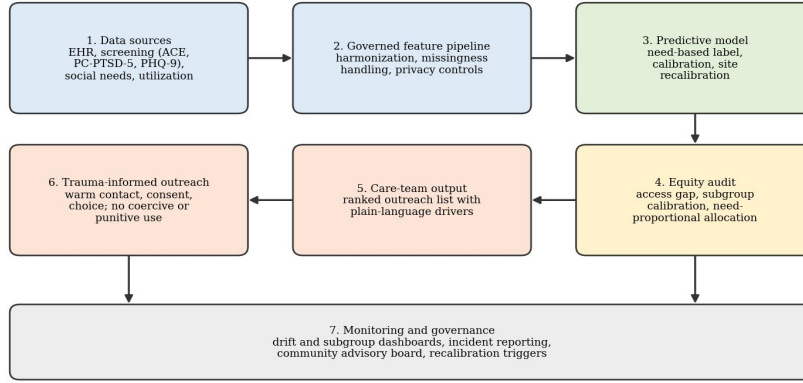


Figure 4: Proposed implementation architecture for AI-driven predictive modeling in trauma-informed integrated care, from data sources and governed feature pipeline to the predictive model, equity audit, care-team output, trauma-informed outreach, and continuous monitoring and governance.

D. Outcome Definition and the Label Problem

The prediction target is a trauma-related crisis within 90 days, defined as an acute behavioral-health emergency or crisis-service contact, or a clinician-documented acute distress episode requiring same-day intervention. Two versions of this outcome matter. The true-need outcome Y^* is whether such a crisis actually occurred. The recorded outcome is whether it was recorded in the health system. The relationship between the two depends on access to care, because patients with limited access are less likely to present to services that generate a record. Letting $A = 1$ denote limited access, the recording probability is

$$P(Y_{obs} = 1 | Y^* = 1, A = a) = r_a, \quad r_0 > r_1 \quad (1)$$

so a model trained on the recorded outcome learns a distorted target. Risk is modeled with a gradient-boosted tree ensemble, whose output is an estimated probability of the form

$$P(Y^* = 1 | x) = \frac{1}{1 + e^{-f(x)}} \quad (2)$$

where $f(x)$ is the additive score produced by the model. The logistic form is used because it yields calibrated probabilities that can be audited; a penalized logistic regression was fitted as a comparator.

E. Implementation Evaluation Plan

For the stepped-wedge study, outcomes are organized by RE-AIM [26] and implementation outcome categories [32] (Table 2). The primary effectiveness outcome is the rate of trauma-related crisis events at 90 days among patients flagged by the model, compared between periods of usual care and AI-supported outreach, analyzed with models that account for clustering by site and time [28]. The primary implementation outcome is reach among patients with limited access to care.

TABLE 2. Planned implementation evaluation using RE-AIM (pre-specified; not yet conducted)

RE-AIM dimension	Planned measure	Planned data source
Reach	Share of true-need patients contacted, overall and by access, deprivation, language, and insurance stratum	EHR; outreach logs; adjudicated sample
Effectiveness	90-day crisis events; engagement in care; patient-reported safety and trust	EHR; patient surveys
Adoption	Share of eligible clinics and clinicians using the outreach list	Site records; staff interviews
Implementation	Fidelity to the trauma-informed protocol; model alert burden; acceptability and feasibility	Audits; surveys
Maintenance	Sustained use at 12 months; recalibration frequency; cost	Site records; finance data

F. Ethics, Privacy, and Governance

A real study would require institutional review board approval, a data use agreement across sites, compliance with applicable privacy law, data minimization, and role-based access. Risk scores would be visible only to the treating team and the outreach team and would not be shared with employers, insurers, or law enforcement. A community advisory board with patient representation would review model purpose, outreach scripts, and audit results, and patients would be able to opt out.

Documentation would follow reporting guidance for prediction models [33], and risk of bias would be appraised with PROBAST [34]. The synthetic validation reported below used no patient data and required no ethical approval.

Technical Validation Setup

A. Synthetic Multicenter Cohort

A synthetic cohort of 16,000 adult patients was generated across six simulated urban sites (A to F, 2,000 to 3,200 patients each) with a fixed random seed. Site-level differences were introduced in mean area deprivation and baseline risk. Each patient was assigned demographic variables, area deprivation (a percentile from 1 to 100), limited English proficiency, uninsured status, housing instability, food insecurity, substance use, chronic condition count, social support, ACE count, PC-PTSD-5 score, PHQ-9 score, and prior-year utilization (emergency visits, outpatient visits, missed appointments). Limited access to care was generated as a function of deprivation, uninsured status, and limited English proficiency. True need Y^* was generated from a logistic function of the clinical and social variables with site-specific intercepts, giving an overall prevalence of 14.7%. Utilization depended on both need and access: patients with limited access had fewer outpatient visits and more missed appointments. The recorded outcome retained a true crisis with probability 0.88 for patients with adequate access and 0.52 for those with limited access (Eq. 1), and screening scores were missing for 10% and 35% of patients, respectively. An adjudicated label, representing structured outreach and clinical review, matched true need with 95% sensitivity and 99% specificity regardless of access. Table 3 gives the characteristics of the simulated sites. All parameters are assumptions chosen to reflect the mechanisms described in the literature; they are not estimates from real patients.

B. Models and Configurations

Three configurations were compared. C1 trained a gradient-boosted model (scikit-learn [35]) on the recorded outcome, the common situation in which the label is whatever the health system captured. C2 trained the same model on the adjudicated need label. C3 used the C2 predictions but allocated outreach capacity across area-deprivation tertiles in proportion to expected need,

$$k_g = K \times \frac{\sum_{i \in g} \hat{p}_i}{\sum_i \hat{p}_i} \quad (3)$$

where K is the total number of patients that can be contacted and $\hat{p}$ is the predicted probability. Outreach capacity was fixed at 10% of the patient panel. Performance was estimated with five-fold cross-validation stratified by site and need, so that all reported metrics are out-of-fold. Generalization to new sites was assessed by leave-one-site-out validation of C2. Confidence intervals were obtained by bootstrap resampling (300 replicates). Model-development sample size and calibration followed recommendations for prediction models [36], [37].

C. Equity and Performance Metrics

Because outreach is limited, the principal metric is reach, the proportion of true-need patients who are flagged. For group g ,

$$Reach_g = \frac{\text{Number of true-need patients flagged in group } g}{\text{Number of true-need patients in group } g} \quad (4)$$

which is the true positive rate used in the equality-of-opportunity criterion [21]. The access gap is the difference in reach between patients with adequate and limited access,

$$\Delta = Reach_{A=0} - Reach_{A=1} \quad (5)$$

and a gap near zero means that patients with equal need have equal chances of being contacted. Calibration by group was assessed by comparing mean predicted risk with the observed true-need rate and by the calibration model

$$\text{logit } P(Y^* = 1) = \alpha + \gamma \cdot \text{logit}(\hat{p}) \quad (6)$$

in which $\alpha = 0$ and $\gamma = 1$ indicate perfect calibration [37]. Workload was summarized by the number needed to contact,

$$NNC = \frac{1}{PPV}, \quad PPV = \frac{TP}{TP + FP} \quad (7)$$

and clinical utility by net benefit at threshold probability p_t [38],

$$NB(p_t) = \frac{TP}{N} - \frac{FP}{N} \times \frac{p_t}{1-p_t} \quad (8)$$

together with the area under the ROC curve (AUC) and the Brier score,

$$Brier = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2 \quad (9)$$

Ground truth for all equity and performance metrics was the true need Y^* , which is known only because the data are synthetic. In practice, an adjudicated sample or structured outreach would be required to approximate it.

Results

A. Cohort Characteristics

Table 3 summarizes the six simulated sites. True need prevalence was 14.7% overall and ranged from 9.3% (Site D) to 20.9% (Site F), tracking mean deprivation. Recorded crisis prevalence was lower, at 10.9% overall. The recording gap was not uniform. Patients with limited access had a higher true prevalence than patients with adequate access (16.5% versus 13.8%) but a lower recorded prevalence (8.5% versus 12.2%), so the recorded data ranked the groups in the opposite order to their true need. True prevalence rose with area deprivation, from 9.9% in the lowest tertile to 13.8% and 20.5% in the middle and highest tertiles, and the share of patients with limited access rose from 24.3% to 45.1%.

TABLE 3. Characteristics of the six simulated sites (synthetic data)

Site	N	True need (%)	Recorded (%)	Limited access (%)	Screening missing (%)	Mean deprivation
Site A	3200	15.8	11.9	34.7	18.8	71
Site B	2600	11.4	8.7	31.0	17.7	65
Site C	3000	17.6	12.7	36.6	19.4	76
Site D	2400	9.3	7.0	30.8	17.5	59
Site E	2800	13.6	10.2	34.0	18.8	69
Site F	2000	20.9	15.2	41.1	20.5	80
All	16,000	14.7	10.9	34.5	18.7	70

B. Overall Predictive Performance

Aggregate performance was almost identical across configurations (Table 4). AUC against true need was 0.827 for C1 and 0.832 for both C2 and C3, and a penalized logistic regression trained on the adjudicated label reached 0.837. When C1 was evaluated against the recorded outcome on which it had been trained, its AUC was 0.825, essentially the same as against true need. Conventional discrimination metrics therefore gave no sign of any problem with the utilization-trained model. At the 10% capacity, overall reach was 40.6% for C1, 42.3% for C2, and 42.0% for C3. Positive predictive value was about 60% (59.8%, 62.3%, and 61.9%), corresponding to roughly 1.6 contacts per true-need patient reached; this high value reflects the combination of a 14.7% prevalence and a capacity set at 10% of the panel.

TABLE 4. Overall performance at 10% outreach capacity (out-of-fold, synthetic data; 95% bootstrap CI in parentheses)

Configuration	AUC (true need)	Brier score	Overall reach (%)	PPV (%)
C1: Utilization label	0.827	0.097	40.6 (38.6–42.3)	59.8
C2: Need-based label	0.832	0.092	42.3 (40.1–44.1)	62.3
C3: C2 + need-proportional allocation	0.832	0.092	42.0 (40.0–43.9)	61.9

C. Equity Audit

The equity audit told a different story (Table 5 and Figure 5). Under C1, 47.1% of true-need patients with adequate access were reached, compared with 30.4% of those with limited access, an access gap

of 16.7 percentage points (95% CI 13.1 to 20.3). Training on the adjudicated label (C2) reduced the gap to 2.3 points (95% CI -1.8 to 6.1), with reach of 43.2% and 40.9% respectively, and C3 produced 43.8% and 39.3% (gap 4.6 points; 95% CI 0.7 to 7.9). The corrective effect of the label was therefore large and the additional effect of reallocation by deprivation was neutral for the access gap.

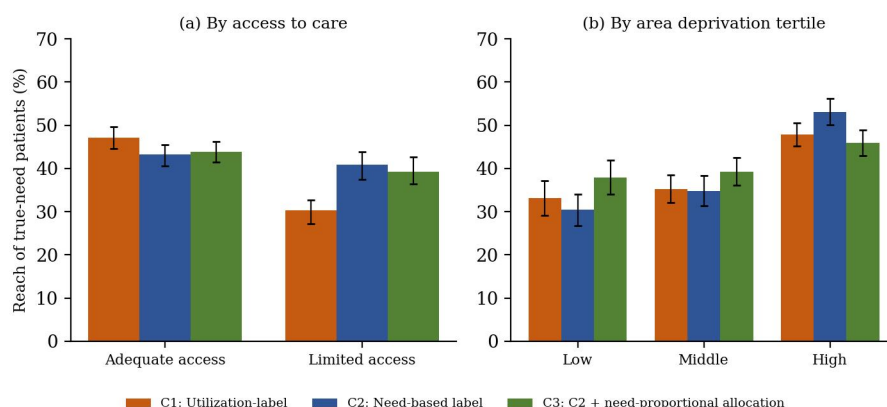


Figure 5: Reach of true-need patients at 10% outreach capacity for the three configurations (a) by access to care and (b) by area-deprivation tertile. Error bars show 95% bootstrap confidence intervals. Synthetic data.

Calibration by access stratum (Figure 6) identified the mechanism. Under C1, the mean predicted risk for patients with limited access was 8.6%, about half their true-need rate of 16.5%, whereas for patients with adequate access it was 12.0% against a true rate of 13.8%. Under C2, the corresponding values were 14.1% against 16.5% and 15.2% against 13.8%. The model trained on recorded outcomes had thus learned that patients with limited access are lower-risk, which is the opposite of the truth.

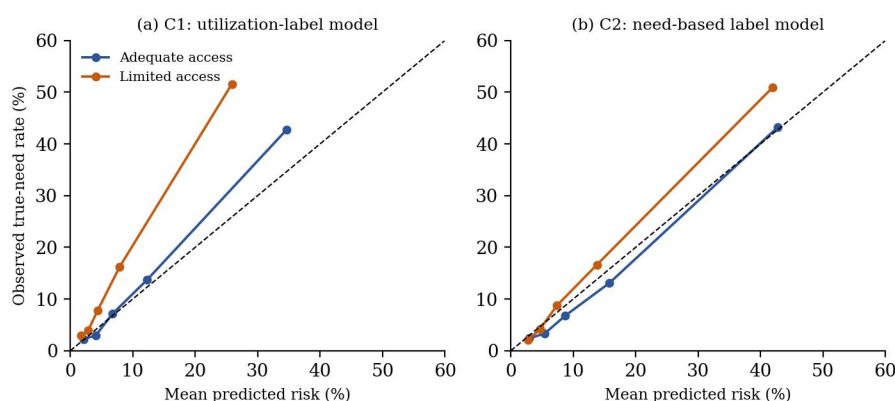


Figure 6: Calibration against true need by access stratum for (a) the utilization-label model C1 and (b) the need-based label model C2. Points show mean predicted risk and observed true-need rate in quintiles of predicted risk within each stratum. The dashed line indicates perfect calibration. Synthetic data.

TABLE 5. Reach of true-need patients by access and deprivation stratum at 10% capacity (synthetic data)

Configuration	Adequate access (%)	Limited access (%)	Access gap, points (95% CI)	Low depr. (%)	High depr. (%)	Depr. gap, points
C1	47.1	30.4	16.7 (13.1 to 20.3)	33.1	47.8	-14.7
C2	43.2	40.9	2.3 (-1.8 to 6.1)	30.5	53.1	-22.6
C3	43.8	39.3	4.6 (0.7 to 7.9)	37.9	45.9	-8.0

The deprivation axis showed a different pattern. Because true need rose steeply with deprivation, a model that ranks patients by risk concentrates outreach in the highest-deprivation tertile. Under C2, reach was 53.1% in the highest tertile and 30.5% in the lowest, a difference of 22.6 points. This is not

unfair in the sense of the access gap, since it follows real differences in need, but it leaves roughly seven in ten true-need patients in the lowest-deprivation tertile unreached. Need-proportional allocation (C3) raised reach in the lowest tertile to 37.9% and lowered it in the highest to 45.9%, narrowing the difference to 8.0 points with almost no change in overall reach (42.3% to 42.0%). Differences by limited English proficiency and insurance status were small under all configurations (absolute gaps 5.4 points or less).

D. Drivers of Predictions

Permutation importance for C2 in one test fold showed that the three utilization features dominated: emergency visits reduced AUC by 0.102 when permuted, outpatient visits by 0.062, and missed appointments by 0.044. Screening and social variables followed at a much lower level: the PC-PTSD-5 score (0.009), area deprivation (0.009), the PHQ-9 score (0.007), ACE count (0.004), and housing instability (0.004). In this simulation the utilization variables were informative by construction, so the ranking should not be read as a finding about real clinics. It nevertheless shows a general hazard: a model that relies heavily on encounter-based features is sensitive to changes in access.

E. Generalization Across Sites

In leave-one-site-out validation of C2 (Figure 7), AUC ranged from 0.812 (Site B) to 0.844 (Site C). Reach at 10% capacity ranged from 35.3% (Site F) to 47.3% (Site D). Calibration-in-the-large varied more: intercepts ranged from -0.22 (Site D) to $+0.49$ (Site F), meaning that a model developed elsewhere over-predicted risk at the lowest-prevalence sites and under-predicted it at the highest-prevalence sites. Discrimination therefore transferred well, whereas risk estimates did not, which supports recalibration for each site before deployment.

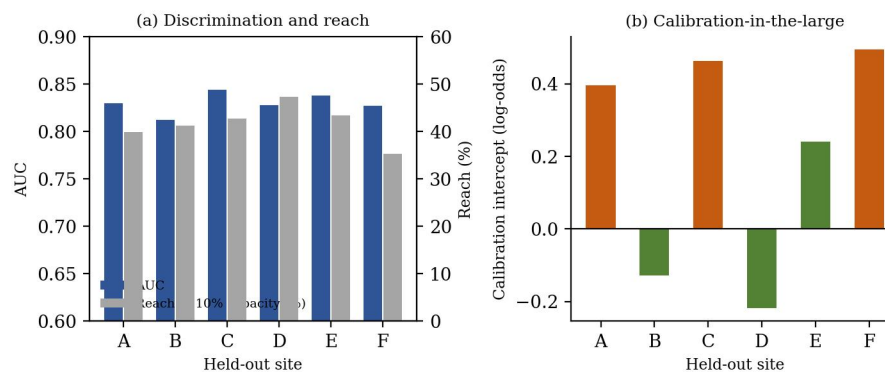


Figure 7: Leave-one-site-out validation of the need-based label model C2: (a) AUC and reach at 10% capacity for each held-out site and (b) calibration intercept for each held-out site. Synthetic data.

F. Clinical Utility

Net benefit was positive for both models and at least as high as the treat-all strategy at every threshold examined, with the largest advantage at higher thresholds. At a threshold probability of 10%, net benefit was 0.0889 for C2, 0.0857 for C1, and 0.0526 for treating everyone; at 20% it was 0.0648, 0.0588, and -0.0659 . The aggregate advantage of C2 over C1 was small because the harm of the utilization label is distributional, concentrated in the group with limited access.

Discussion

The technical validation yields four lessons that should shape any real implementation. First, the label is the most consequential modeling decision. Three configurations with nearly identical AUC (0.827 to 0.832) differed by 14.4 points in the access gap. Training on recorded crisis contacts made the model blind to need among people with limited access, and no conventional aggregate metric flagged the problem. This reproduces, in a trauma-informed setting, the mechanism described for cost-based risk scores [6], and it follows from a plain fact about underserved settings: people who face barriers to care leave fewer traces in the record. Teams should therefore invest in need-based labels, such as adjudicated samples or structured outreach follow-up.

Second, equity is not a single quantity. The access gap and the deprivation difference measure different things, and a configuration that improved one did not necessarily improve the other. Reallocating outreach in proportion to expected need raised reach among true-need patients in the

lowest-deprivation tertile at the cost of reach in the highest. Whether that is the right trade-off is an ethical and organizational decision, not a technical one, and it is exactly the kind of decision that a community advisory board should help to make.

Third, deployment across sites requires local calibration. Discrimination was stable, but calibration intercepts spanned about 0.7 log-odds units. A single pooled model would have over-flagged low-prevalence sites and under-flagged high-prevalence sites if thresholds had been applied to predicted probability. This echoes the experience of widely implemented clinical models that performed worse in new settings than in development [16], and it supports the monitoring and recalibration layer of the proposed architecture. Federated approaches can enlarge training data without pooling records, but they do not remove the need for local calibration [30], [37].

Fourth, the technical layer is only part of the intervention. The main benefit of a predictive list lies in the quality of the contact that follows. A well-calibrated, fair model used to deliver intrusive or poorly timed outreach could harm patients, which is why the framework couples every modeling choice to a trauma-informed requirement and an audit check (Table 1). Downstream personalization, such as the functions in Figure 3, raises the same questions and should follow an audit of the risk-prediction layer.

This study has important limitations. The empirical results come from a synthetic cohort whose mechanisms and parameters were assumptions. They show how the pipeline behaves when label bias of a specified kind is present; they do not estimate its size in real clinics, and different assumptions would give different numbers. The adjudicated label assumed near-perfect ascertainment, and only access, deprivation, language, and insurance were audited; a real audit would also examine race, ethnicity, sex, age, and disability. The models predict risk and do not estimate the effect of outreach, which requires the stepped-wedge evaluation described in Section IV. The sample of six sites is small for studying between-site variation. Future work should replace the synthetic cohort with governed data from partner sites, validate need labels, evaluate patient experience of outreach, test federated training, and report the implementation outcomes in Table 2.

Conclusion

This paper set out a multicenter implementation framework for using AI-driven predictive modeling in trauma-informed integrated care and tested its predictive and equity-audit components on a synthetic six-site cohort. The framework links the six principles of a trauma-informed approach to specific design requirements and audit checks, and specifies a hybrid stepped-wedge evaluation based on CFIR and RE-AIM. The technical validation showed that models with almost identical AUC can differ greatly in whom they reach: a utilization-trained model reached 30.4% of true-need patients with limited access against 47.1% of those with adequate access, whereas training on an adjudicated need label reduced the gap to 2.3 points. It also showed that discrimination transfers across sites more readily than calibration, and that allocation rules involve value-laden trade-offs between groups.

The central implication is that the safety of AI in trauma-informed care depends on how labels are chosen, how equity is measured, how models are recalibrated, and how outreach is delivered, more than on the choice of algorithm. The authors recommend that the next step be the real multicenter study specified here, with governed data, need-based labels, community oversight, and pre-registered implementation outcomes, so that the promise of AI to support whole-person care in underserved urban settings is tested without widening the disparities it aims to reduce.

References

- [1] Substance Abuse and Mental Health Services Administration. (2014). SAMHSA's concept of trauma and guidance for a trauma-informed approach (HHS Publication No. SMA 14-4884). Rockville, MD: SAMHSA.
- [2] Felitti, V. J., Anda, R. F., Nordenberg, D., et al. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the Adverse Childhood Experiences (ACE) Study. *American Journal of Preventive Medicine*, 14(4), 245–258.
- [3] Hughes, K., Bellis, M. A., Hardcastle, K. A., et al. (2017). The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *The Lancet Public Health*, 2(8), e356–e366.
- [4] Roberts, A. L., Gilman, S. E., Breslau, J., Breslau, N., & Koenen, K. C. (2011). Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the United States. *Psychological Medicine*, 41(1), 71–83.
- [5] Braveman, P., & Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. *Public Health Reports*, 129(Suppl 2), 19–31.

- [6] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- [7] Purtle, J. (2020). Systematic review of evaluations of trauma-informed organizational interventions that include staff trainings. *Trauma, Violence, & Abuse*, 21(4), 725–740.
- [8] Valentijn, P. P., Schepman, S. M., Opheij, W., & Bruijnzeels, M. A. (2013). Understanding integrated care: a comprehensive conceptual framework based on the integrative functions of primary care. *International Journal of Integrated Care*, 13, e010.
- [9] Archer, J., Bower, P., Gilbody, S., et al. (2012). Collaborative care for depression and anxiety problems. *Cochrane Database of Systematic Reviews*, (10), CD006525.
- [10] Katon, W. J., Lin, E. H. B., Von Korff, M., et al. (2010). Collaborative care for patients with depression and chronic illnesses. *New England Journal of Medicine*, 363(27), 2611–2620.
- [11] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613.
- [12] Prins, A., Bovin, M. J., Smolenski, D. J., et al. (2016). The Primary Care PTSD Screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. *Journal of General Internal Medicine*, 31(10), 1206–1211.
- [13] Billioux, A., Verlander, K., Anthony, S., & Alley, D. (2017). Standardized screening for health-related social needs in clinical settings: the Accountable Health Communities Screening Tool. *NAM Perspectives*. Washington, DC: National Academy of Medicine.
- [14] Simon, G. E., Johnson, E., Lawrence, J. M., et al. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175(10), 951–960.
- [15] Belsher, B. E., Smolenski, D. J., Pruitt, L. D., et al. (2019). Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry*, 76(6), 642–651.
- [16] Wong, A., Otles, E., Donnelly, J. P., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065–1070.
- [17] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 195.
- [18] Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1547.
- [19] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 115.
- [20] Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), 2176–2182.
- [21] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 29.
- [22] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866–872.
- [23] Wiens, J., Saria, S., Sendak, M., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340.
- [24] World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: WHO.
- [25] Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science*, 4, 50.
- [26] Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health*, 89(9), 1322–1327.
- [27] Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M., & Stetler, C. (2012). Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical Care*, 50(3), 217–226.
- [28] Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350, h391.
- [29] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282.
- [30] Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.

- [31] Tervalon, M., & Murray-García, J. (1998). Cultural humility versus cultural competence: a critical distinction in defining physician training outcomes in multicultural education. *Journal of Health Care for the Poor and Underserved*, 9(2), 117–125.
- [32] Proctor, E., Silmere, H., Raghavan, R., et al. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health*, 38(2), 65–76.
- [33] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378.
- [34] Wolff, R. F., Moons, K. G. M., Riley, R. D., et al. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1), 51–58.
- [35] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [36] Riley, R. D., Ensor, J., Snell, K. I. E., et al. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, m441.
- [37] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17, 230.
- [38] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.