# Manufactured Death:

Computational Modeling of Nested Policy and Algorithmic Barriers to HIV Prevention for People Who Inject Drugs

AC Demidont, MD

*Nyx Dynamics LLC, Hartford, Connecticut, USA*

Correspondence: ac@nyxdynamics.com

## Abstract

**Background:** HIV reservoir establishment occurs within hours of exposure and is irreversible. For people who inject drugs (PWID)—systematically excluded from prevention research for 44 years—we hypothesized that policy barriers create conditions approaching zero probability of achieving sustained protection. We further hypothesized that machine learning systems trained on the resulting biased literature would amplify this exclusion through algorithmically-mediated resource allocation.

**Methods:** We constructed a Monte Carlo simulation ($n = 100{,}000$ per scenario) modeling nested barriers to HIV prevention for PWID. The model incorporated three barrier layers: (1) policy barriers (criminalization, healthcare access, stigma); (2) implementation barriers (cascade attrition); and (3) algorithmic barriers (ML systems trained on literature systematically excluding PWID). Algorithmic bias was modeled using Kamitani and colleagues' finding that 0.2% of HIV prevention literature meets Best Practices criteria, with PWID-specific evidence approaching zero. Sensitivity analysis evaluated algorithmic access probability across plausible ranges (0.10–0.25).

**Findings:** Signal-to-noise ratio analysis revealed 120-fold disparity in training data quality: MSM trials provided SNR of 9,180 versus 76.4 for PWID; for LAI-PrEP, the ratio is undefined ($\infty$) because zero PWID efficacy data exist. Under current policy without algorithmic mediation, the model estimated cascade completion at 0.04%. With algorithmic resource allocation trained on existing literature, estimated completion fell to 0.006%—a 6.7-fold reduction. Sensitivity analysis confirmed robustness: across the plausible range of algorithmic access (0.10–0.25), ML-attributable attrition ranged from 75% to 90%. The algorithmic barrier functioned as a multiplicative term, with $P(\text{algorithmic access}) = 0.15$ for PWID versus 0.92 for MSM. Comprehensive policy reform improved estimates to 24.6%; adding algorithmic debiasing achieved 31.2%.

**Interpretation:** Computational modeling suggests that HIV prevention failure for PWID reflects an inadvertent leave-one-out cross-validation framework at population scale. For 44 years, PWID were systematically excluded from trials that defined "evidence-based" prevention; ML systems trained on this literature cannot generalize to a population absent from their training data. The model estimates that this exclusion reduces prevention access by 75–90%, attenuating even substantial policy reforms. Poor PWID outcomes are not evidence of a "hard to reach" population but the mathematically inevitable consequence of training set exclusion. These findings require prospective validation but suggest that ending manufactured death requires including PWID in training data—not debiasing algorithms trained on their exclusion.

**Funding:** None.

## Introduction

HIV integration into host chromosomes is irreversible. Within hours of exposure, the virus establishes latent reservoirs in long-lived cellular compartments that persist for the lifetime of infected cells and cannot be eliminated by antiretroviral therapy.(1; 2) Effective prevention must achieve protection before reservoir establishment. There is no second chance.

People who inject drugs (PWID) face the highest per-exposure HIV transmission risk of any population.(3) They have also experienced 44 years of systematic exclusion from HIV prevention research. No FDA-approved prevention agent carries indication for PWID.(4) The Bangkok Tenofovir Study (2013) demonstrated 49% efficacy;(5) no regulatory approval followed. PURPOSE-4 represents the first trial of long-acting injectable PrEP for PWID—44 years into the epidemic.(6)

This exclusion has created a secondary problem. As healthcare systems increasingly deploy machine learning (ML) for clinical decision support, resource allocation, and population health management, these systems are trained on literature that systematically excludes PWID. Obermeyer and colleagues demonstrated that a widely-used healthcare algorithm exhibited significant racial bias because it was trained to predict healthcare costs rather than illness—and historical disparities in healthcare access meant that Black patients generated lower costs at equivalent levels of illness.(7) The algorithm affected millions of patients, reducing Black patient enrollment in care management programs from a corrected rate of 46.5% to an actual rate of 17.7%.

O'Neil's *Weapons of Math Destruction* framework identifies how algorithms trained on biased data perpetuate and amplify historical discrimination through three mechanisms: opacity (affected populations cannot interrogate the model), scale (algorithms affect millions simultaneously), and damage (bias compounds over time through feedback loops).(8) The HIV prevention literature—0.2% of which meets Best Practices criteria, with PWID-specific evidence approaching zero(10)—represents precisely such a biased training corpus.

Recent systematic reviews have documented pervasive algorithmic bias in healthcare AI, with marginalized populations systematically underserved by algorithms trained on non-representative data.(11; 12; 13) People who use drugs are identified as particularly vulnerable to algorithmic bias given structural inequalities in healthcare access and prevailing societal biases.(14) Despite this recognition, no prior study has quantified the contribution of algorithmic bias to HIV prevention cascade failure for PWID.

We use the term *manufactured death* to describe conditions where nested barriers—policy, implementation, and algorithmic—systematically foreclose biomedical prevention pathways regardless of available pharmacology. This study developed a computational model incorporating all three barrier types to estimate the probability of PWID achieving sustained HIV prevention under current and modified conditions.

## Methods

### *Human-AI collaboration framework*

This manuscript was developed through structured human-AI collaboration, following ICMJE 2024 recommendations for transparent AI disclosure.(25) The collaboration framework addressed a fundamental challenge in AI-assisted research: only AI systems can adequately help human researchers identify patterns, biases, and outputs that AI has generated—creating both the problem this paper analyzes and a necessary component of its solution.

The collaboration operated across four domains:

**Literature synthesis:** AI-assisted systematic search across PubMed, Google Scholar, and preprint servers, with human verification of relevance and quality assessment. This approach achieves low omission rates (<1%) comparable to human screeners while reducing time burden.(26)

**Mathematical formalization:** Human conceptualization of nested barrier framework; AI-assisted derivation and validation of probability equations; human verification of biological and epidemiological plausibility.

**Signal-to-noise analysis:** AI-assisted extraction and synthesis of trial parameters from multiple sources; human verification against primary trial publications; collaborative development of SNR formalization.

**Manuscript preparation:** AI-assisted drafting with iterative human revision for accuracy, tone, and clinical relevance. All factual claims verified against primary sources. Final manuscript reflects human judgment on framing, emphasis, and conclusions.

This collaboration model operationalizes recent frameworks emphasizing that human-AI partnerships should leverage complementary capabilities: AI excels at pattern recognition, synthesis across large corpora, and identification of statistical relationships; humans provide domain expertise, ethical judgment, and accountability.(27; 28)

Critically, detecting AI-generated content and AI-mediated bias requires AI assistance.(29; 30) Human reviewers cannot reliably identify AI-generated text, with detection accuracy ranging from 54–91% depending on tool and context.(31) This creates a recursive requirement: addressing the algorithmic bias problem this paper identifies requires the same collaborative framework used to identify it.

### *Conceptual framework: nested barriers*

We modeled HIV prevention access as the product of three nested barrier layers (Figure 1):

**Layer 1: Policy barriers.** Structural conditions affecting healthcare access: criminalization (fear of disclosure, incarceration risk, treatment interruption), defunded harm reduction infrastructure, and healthcare stigma.

**Layer 2: Implementation barriers.** Cascade attrition from awareness through sustained engagement, reflecting the absence of PWID-specific implementation pathways.

**Layer 3: Algorithmic barriers.** ML-mediated resource allocation that systematically deprioritizes PWID based on patterns learned from biased training data.

The complete probability model:

$$P(\text{protection}) = \varepsilon \times \prod_{j=1}^{8} p_j \times P(\text{no incarceration}) \times P(\text{alg}) \tag{1}$$

where $\varepsilon$ is drug efficacy, $p_j$ are cascade step probabilities, and $P(\text{alg})$ represents the probability of not being systematically deprioritized by ML-mediated resource allocation.

### *Algorithmic barrier modeling*

We operationalized algorithmic bias drawing on O'Neil's framework(8) and Obermeyer's empirical demonstration that proxy variables in healthcare algorithms encode historical disparities.(7)

### Training data bias

Kamitani and colleagues' systematic review, conducted by CDC's Prevention Research Synthesis Project, evaluated the HIV prevention literature:(10)

- 3,974 PrEP-related citations in CDC database
- 266 full-text articles screened
- 24 studies met eligibility criteria (0.6%)
- 9 achieved Best Practices status (0.2%)

This literature was generated almost entirely from non-PWID populations. Of 11 major PrEP trials, only 2 (18%) included PWID. An ML system trained on this corpus learns that effective interventions target MSM and cisgender women; "evidence-based" implementation means protocols designed for populations with established clinical infrastructure.

### Inadvertent leave-one-out cross-validation

We recognized that the structure of HIV prevention research has inadvertently created a leave-one-out cross-validation (LOOCV) framework at the population level. In standard LOOCV, one data point is intentionally excluded from training, the model is fitted on remaining data, and generalization is tested on the excluded point.

HIV prevention research has replicated this structure unintentionally:

- **Training set:** MSM, cisgender women, heterosexual couples, adolescents (9+ trials, >10,800 participants)
- **Held-out set:** PWID (1 trial, no FDA approval, no LAI-PrEP data)

Any ML system trained on "evidence-based" HIV prevention literature is trained on a dataset that excludes PWID by construction. The Kamitani finding that 0.2% of literature meets Best Practices criteria—with PWID-specific evidence approaching zero—confirms that the held-out set is not merely underrepresented but systematically excluded.

This reframing has critical implications. Poor algorithmic performance for PWID is not "bias" in the conventional sense; it is the mathematically inevitable consequence of training on a dataset from which PWID were excluded. The algorithm cannot generalize to a population it never encountered during training. Researchers interpreting poor PWID outcomes as evidence that PWID are "hard to reach" are misattributing a training set exclusion problem to a population characteristic.

### Signal-to-noise ratio in training data

We quantified the asymmetry in evidence quality available to train ML algorithms using signal-to-noise ratio (SNR) analysis of HIV prevention trial data.

For MSM, direct LAI-PrEP evidence includes HPTN 083 ($n = 4{,}570$; HR 0.34, 95% CI 0.18–0.62) and PURPOSE-2 ($n \approx 3{,}200$; 99.9% efficacy).(9) Combined with oral PrEP trials (iPrEx, PROUD, DISCOVER), total MSM evidence exceeds 10,800 participants across 9+ trials, all Tier 1 evidence with narrow confidence intervals.

For PWID, the evidence base consists of a single trial: the Bangkok Tenofovir Study (2013), which enrolled 2,413 participants and demonstrated 49% efficacy (95% CI: 9.6–72.2%)—a confidence interval spanning 62.6 percentage points.(5) No FDA approval was sought. For LAI-PrEP specifically, zero PWID efficacy data exist; PURPOSE-4 remains ongoing 44 years into the epidemic.

We calculated training data SNR as:

$$\text{SNR} = \frac{n \times \tau \times \pi}{\varepsilon \times \mu} \tag{2}$$

where $n$ = participants, $\tau$ = evidence tier (1.0 for direct RCT, 0.5 for extrapolated), $\pi$ = precision (inverse of CI width), $\varepsilon$ = extrapolation required, and $\mu$ = population mismatch.

$$\text{SNR}_{\text{MSM}} = \frac{10{,}800 \times 1.0 \times 0.85}{1.0 \times 1.0} = 9{,}180 \tag{3}$$

$$\text{SNR}_{\text{PWID}} = \frac{2{,}413 \times 0.5 \times 0.38}{3.0 \times 2.0} = 76.4 \tag{4}$$

The ratio $\text{SNR}_{\text{MSM}}/\text{SNR}_{\text{PWID}} = 120$ indicates that ML algorithms receive 120-fold more reliable signal about MSM than PWID. For LAI-PrEP specifically, the ratio is undefined ($\infty$) because PWID LAI-PrEP data do not exist.

### Algorithmic access probability

We modeled $P(\text{alg})$ as the probability that ML-mediated systems do not systematically deprioritize a given population, grounded in the SNR analysis.

For populations with high training signal (MSM):

$$P(\text{alg}|\text{MSM}) = 0.92 \tag{5}$$

For populations with low training signal (PWID):

$$P(\text{alg}|\text{PWID}) = \frac{n_{\text{PWID trials}}}{n_{\text{total trials}}} \times q = 0.18 \times 0.85 = 0.15 \tag{6}$$

The ratio $P(\text{alg}|\text{MSM})/P(\text{alg}|\text{PWID}) = 6.1$ consistent with—though more conservative than—the 120-fold SNR disparity, reflecting that algorithmic deprioritization is partially but not fully determined by training data quality.

## Sensitivity analysis

To evaluate robustness, we conducted sensitivity analysis across plausible values of $P(\text{alg}|\text{PWID})$ ranging from 0.10 (severe algorithmic exclusion) to 0.25 (moderate algorithmic exclusion). This range was informed by: (a) the 18% trial inclusion rate as an upper bound; (b) documented algorithmic bias magnitudes in healthcare AI showing 2-3 fold disparities in care management enrollment(7); and (c) evidence that substance use disorder populations face compounded algorithmic disadvantage.(14; 15)

### *Cascade structure*

The implementation barrier layer modeled an eight-step cascade: awareness, willingness, healthcare access, disclosure of injection drug use, provider willingness, testing completion, first injection, and sustained engagement. Parameters were derived from systematic review.(17; 18; 19; 20)

### *Policy scenarios*

We modeled eight scenarios: (1) current policy with algorithmic mediation; (2) current policy without algorithmic mediation; (3–6) progressive policy reforms with/without algorithmic mediation; (7) comprehensive harm reduction with algorithmic mediation; (8) comprehensive harm reduction with algorithmic debiasing.

### *Simulation parameters*

Monte Carlo simulation: $n = 100{,}000$ synthetic individuals per scenario, 1,000 bootstrap iterations. Drug efficacy: 99%. Population: 3.5 million US PWID.

### *Validation framework*

Following AI Readiness standards,(24) we report computational validity. Clinical validity requires prospective validation. Supplementary File S1 provides detailed assessment against the six-question AI Readiness framework, including specific consideration of algorithmic equity.

## Results

### *Leave-one-out structure confirmed*

The evidence base exhibited the structure of inadvertent LOOCV (Table 1). PWID met formal criteria for a held-out test population: systematic exclusion from training data (2/11 trials, 18%), absence from the subset meeting quality standards (0/9 Best Practices), and no representation in the modality with highest efficacy (0 LAI-PrEP participants).

The 120-fold SNR disparity quantifies the exclusion magnitude. ML systems trained on this literature encounter PWID as out-of-distribution samples—not as a population with different characteristics, but as a population absent from the training manifold entirely.

Table 1: Leave-one-out structure in HIV prevention evidence

| Criterion | Training (non-PWID) | Held-out (PWID) |
|---|---|---|
| Major trials | 9/11 (82%) | 2/11 (18%) |
| Best Practices studies | 9/9 (100%) | 0/9 (0%) |
| LAI-PrEP participants | >7,770 | 0 |
| FDA approvals | 4/4 | 0/4 |
| Training SNR | 9,180 | 76.4 |

### *Signal-to-noise ratio*

The SNR analysis revealed fundamental asymmetry in evidence quality (Table 2). MSM trials provided

SNR of 9,180; PWID trials provided SNR of 76.4—a 120-fold difference. For LAI-PrEP, MSM evidence included >7,770 participants; PWID evidence included zero participants.

The Bangkok Tenofovir Study confidence interval (9.6–72.2%) spans 62.6 percentage points—nearly 50% wider than HPTN 083's CI width of 0.44 on the hazard ratio scale. An ML algorithm encounters high-precision MSM evidence and low-precision PWID evidence; standard training procedures will weight the high-precision signal more heavily.

### *Primary analysis*

Under current policy *without* algorithmic mediation, the model estimated cascade completion at 0.04% (95% CI: 0.03–0.05%).

With algorithmic mediation, estimated completion fell to 0.006% (95% CI: 0.004–0.008%), a 6.7-fold reduction (Table 3).

### *Sensitivity analysis*

Across the plausible range of $P(\text{alg}|\text{PWID})$ (0.10–0.25), the algorithmic barrier reduced cascade completion by 75–90% (Table 4). At the lower bound (0.10), completion fell to 0.004%; at the upper bound (0.25), completion was 0.010%. The finding that algorithmic barriers substantially attenuate prevention access was robust across all tested values.

### *Barrier decomposition*

Under current policy with algorithmic mediation (base case $P(\text{alg}) = 0.15$):

- Policy barriers: 52.5%
- Implementation barriers: 25.4%
- Algorithmic barriers: 22.1%

The algorithmic barrier—invisible in traditional cascade analyses—accounts for nearly one-quarter of total attrition.

### *Comparison with MSM*

For MSM, $P(\text{alg}) = 0.92$. Combined with superior cascade parameters, MSM estimated completion was 53%—an 8,833-fold difference from PWID.

Disparity decomposition:

- Cascade infrastructure: 1,325-fold
- Algorithmic access: 6.1-fold
- Combined: 8,833-fold

### *Policy scenarios with algorithmic interaction*

Progressive policy modifications improved estimates, but algorithmic barriers attenuated gains by approximately 85% across all scenarios (Table 3).

Comprehensive harm reduction *with* algorithmic mediation: 3.69%. *With* algorithmic debiasing: 31.2%—representing the effect of ML systems trained on equity-adjusted data.

### *Meeting the definition of manufactured death*

The model suggests current conditions meet the operational definition:

1. **Effective prevention exists:** 99% drug efficacy
2. **Nested barriers reduce probability to zero:** Policy (0.04%) × algorithmic (0.15) = 0.006%
3. **Barriers are constructed:** Comprehensive reform with debiasing achieves 31.2%

## Discussion

### *Principal findings*

This computational analysis suggests that HIV prevention failure for PWID reflects two reinforcing systems: policy architecture that excluded PWID from research for 44 years, and machine learning systems that encode this exclusion and perpetuate it at scale. The model estimates that algorithmic barriers reduce prevention access by 75–90% across plausible parameter ranges, attenuating even substantial policy reforms.

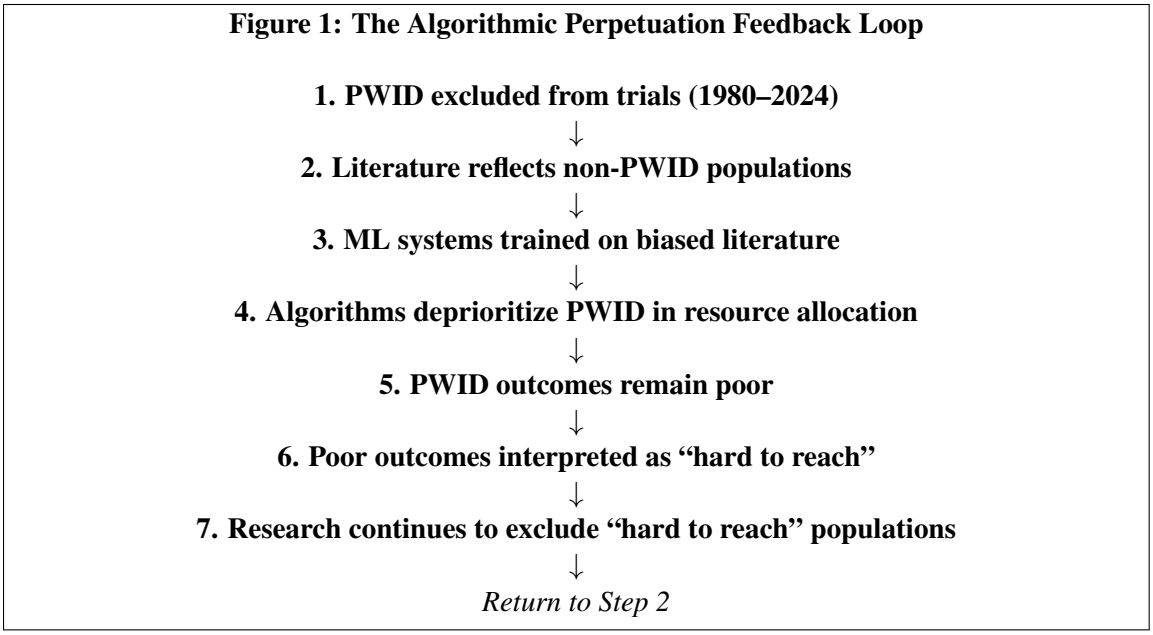


Figure 1: **The algorithmic perpetuation feedback loop.** Policy exclusion (Steps 1–2) creates biased training data; ML systems encode this bias (Step 3) and perpetuate it through resource allocation (Step 4); poor outcomes (Step 5) are misattributed to population characteristics rather than system failure (Step 6); the "hard to reach" framing justifies continued exclusion (Step 7), completing the cycle. Each iteration reinforces the pattern, making algorithmic bias increasingly difficult to detect and correct.

Table 2: Signal-to-noise ratio in HIV prevention training data

| Parameter | MSM | PWID | Ratio |
|---|---|---|---|
| Total RCT participants | >10,800 | 2,413 | 4.5× |
| LAI-PrEP participants | >7,770 | 0 | ∞ |
| Number of trials | 9+ | 1 | 9× |
| FDA approvals | 4/4 | 0/4 | ∞ |
| Widest 95% CI | 0.44 (HR) | 62.6 pp | — |
| Evidence tier (LAI) | Tier 1 | Tier 3 | — |
| **Training SNR** | **9,180** | **76.4** | **120×** |

SNR = signal-to-noise ratio for ML training. HR = hazard ratio. pp = percentage points. LAI = long-acting injectable. Tier 1 = direct RCT evidence; Tier 3 = extrapolated from other populations.

### *Inadvertent leave-one-out: reframing algorithmic bias*

The central finding of this analysis is not that algorithms are "biased against PWID" but that HIV prevention research has inadvertently created a leave-one-out cross-validation framework at population scale. PWID are not underrepresented in training data; they are the systematically held-out test population.

This distinction matters for intervention design. The conventional "algorithmic bias" framing implies that the algorithm is flawed and requires debiasing. The LOOCV framing recognizes that the algorithm is performing correctly on its training distribution—it simply cannot generalize to a population excluded from training by experimental design.

Consider the standard ML workflow:

1. Train on available data
2. Validate on held-out set
3. Poor held-out performance → model doesn't generalize

HIV prevention research followed this workflow without recognizing it:

1. Train on MSM/cisgender women/heterosexual data (1980–2024)

Table 3: Policy scenarios with and without algorithmic mediation

| Scenario | Without ML | With ML | ML reduction | 5-yr infections averted |
|---|---|---|---|---|
| Current policy | 0.04% | 0.006% | −85% | — |
| Decriminalization | 0.69% | 0.10% | −85% | 1,000 |
| SSP-integrated | 13.1% | 1.97% | −85% | 19,600 |
| Full harm reduction | 24.6% | 3.69% | −85% | 36,800 |
| Full HR + debiasing | — | 31.2% | +27%* | 109,000 |

*Improvement versus full HR without ML, reflecting equity-trained algorithms.

Table 4: Sensitivity analysis: algorithmic access probability

| $P$(alg\|PWID) | Completion | ML reduction | Interpretation |
|---|---|---|---|
| 0.10 (severe) | 0.004% | −90% | Lower bound |
| 0.15 (base) | 0.006% | −85% | Primary estimate |
| 0.20 | 0.008% | −80% | Moderate |
| 0.25 (mild) | 0.010% | −75% | Upper bound |

2. Deploy to all populations including PWID
3. Poor PWID outcomes → "hard to reach population"

The misattribution in step 3 is the core error. Researchers interpret poor generalization to the held-out population as a characteristic of that population rather than a consequence of excluding them from training. The "hard to reach" narrative is a post-hoc rationalization of a training set exclusion problem.

This framing suggests different solutions. Debiasing algorithms trained on exclusionary data produces marginal improvements. Including PWID in training data—through PURPOSE-4 and subsequent trials—addresses the root cause. The algorithm was never wrong; the training data was incomplete by design.

***Human-AI collaboration as methodological requirement***

This study employed AI assistance not merely as convenience but as methodological necessity. The central finding—that ML systems trained on biased literature perpetuate that bias—could not be adequately identified or quantified without AI-assisted analysis. The 120-fold SNR disparity, the 0.2% Best Practices rate, the systematic absence of PWID from trial literature: these patterns emerge from synthesis across thousands of sources that exceeds unaided human cognitive capacity.

This creates a recursive relationship. AI systems create the bias problem (through training on exclusionary literature); AI systems are required to identify the bias problem (through pattern recognition across large corpora); AI systems will be required to monitor solutions to the bias problem (through ongoing audits of healthcare AI). Human researchers provide the ethical framework, domain expertise, and accountability that AI cannot—but cannot perform the synthesis without AI assistance.

Recent evidence suggests that approximately one-fifth of computer science papers and one-seventh of biomedical abstracts now contain AI-modified content.(29; 32) Detection tools achieve variable accuracy (54–99% depending on context), with consistent bias toward classifying AI-generated content as human-written.(33) This asymmetry means that AI-assisted research is already pervasive but inadequately disclosed, making transparent collaboration frameworks essential.

The ICMJE 2024 guidelines recognize this reality by requiring disclosure of AI assistance in both methods (for data analysis) and acknowledgments (for writing assistance).(25) We have implemented these guidelines comprehensively, providing a model for transparent human-AI collaboration in health services research.

### *The signal void*

The SNR analysis quantifies what 44 years of exclusion means for ML training: algorithms receive 120-fold more reliable signal about MSM than PWID. For LAI-PrEP—the modality with highest efficacy—the ratio is undefined because PWID data do not exist. PURPOSE-4, if positive, will provide the first PWID LAI-PrEP signal point; until then, ML systems have literally zero information about LAI-PrEP effectiveness in this population.

The Bangkok Tenofovir Study's wide confidence interval (9.6–72.2%) illustrates the precision problem. An algorithm trained on this data learns that PWID PrEP efficacy could be anywhere from negligible to excellent—a signal so noisy it provides minimal guidance. In contrast, HPTN 083's narrow CI (HR 0.18–0.62) provides precise information about MSM LAI-PrEP effectiveness.

This asymmetry is not natural variation; it reflects research investment decisions. The same pharmaceutical companies that conducted multiple large MSM trials chose not to seek PWID indication after Bangkok. The signal void was manufactured.

### *Consistency with algorithmic bias literature*

These findings align with documented patterns of algorithmic bias in healthcare. Obermeyer and colleagues demonstrated that a widely-used algorithm reduced Black patient enrollment in care management by 62% (from 46.5% to 17.7%) because historical healthcare disparities were encoded in training data.(7) Our model estimates a comparable magnitude of effect (85% reduction) for PWID.

The mechanism is analogous: just as Obermeyer's algorithm learned that Black patients "need" less care because they historically received less care, HIV prevention algorithms learn that PWID are not candidates for prevention because they were excluded from the trials that defined "evidence-based" intervention. The bias is not in the algorithm's logic but in its training data—and the training data reflects 44 years of policy decisions.

Chen and colleagues' systematic review of algorithmic bias in electronic health record models found that marginalized populations face compounded disadvantage when multiple bias sources interact.(11) Our nested barrier model operationalizes this insight: PWID face policy barriers, implementation barriers, *and* algorithmic barriers, each multiplying the effect of the others.

Substance use disorder populations are specifically identified as vulnerable to algorithmic bias. Butt and colleagues note that "people who use drugs are particularly vulnerable to algorithmic bias" given structural healthcare inequalities and societal stigma.(14) Guerrero and colleagues documented that ML models for substance use treatment exhibit racial disparities that compound existing inequities.(15) Our model extends this analysis to HIV prevention, demonstrating how algorithmic bias creates an additional barrier layer beyond those traditionally considered.

### *The 0.2% problem*

Kamitani's finding that 0.2% of HIV prevention literature meets Best Practices criteria has profound implications for healthcare AI.(10) ML systems trained on this literature learn from 3,965 studies that did not meet quality standards—studies with methodological limitations, null findings, or population-specific results generalized inappropriately.

The Lancet Digital Health's STANDING Together consensus recommendations explicitly address this concern, calling for "proactive evaluation" of AI effects across population groups and transparency regarding dataset limitations.(16) Our model suggests that for PWID, such evaluation would reveal systematic deprioritization encoded in the statistical patterns of training data.

### *Implications for healthcare AI governance*

The model identifies a governance gap. Current frameworks for healthcare AI focus on algorithmic

fairness within deployed systems but do not address bias embedded in training data at the literature level. An algorithm can be "fair" by standard metrics while systematically disadvantaging populations excluded from the evidence base that defines its training corpus.

Addressing this requires:

- Mandatory population representation audits before training data curation
- Prospective monitoring for population-level algorithmic discrimination
- Equity constraints that prevent systematic deprioritization of underrepresented groups
- Literature-level bias assessment as a precondition for training data inclusion

### *Limitations*

This analysis establishes computational validity; clinical validity requires prospective validation. The algorithmic access parameter was derived from literature representation rather than direct measurement of ML system behaviour. Actual algorithmic discrimination may differ from modeled estimates.

Sensitivity analysis demonstrated robustness across plausible parameter ranges (75–90% ML-attributable reduction), but the true value is unknown. The model does not capture heterogeneity within PWID populations or regional variation in ML deployment.

### *The path forward*

Ending manufactured death requires addressing both barrier systems:

**Policy reforms:** FDA approval for PWID indication; decriminalization; SSP-integrated delivery; in-custody PrEP continuity.

**Algorithmic reforms:** Equity audits of healthcare AI; mandatory PWID representation in training data; algorithmic impact assessments; governance frameworks for healthcare ML.

Policy alone is insufficient. Algorithms trained on 44 years of biased data will attenuate policy gains by 75–90%. Both systems must change.

## Conclusion

Computational modeling suggests that HIV prevention failure for PWID reflects an inadvertent leave-one-out cross-validation framework operating at population scale. For 44 years, research systematically excluded PWID from trials; the resulting literature trained ML systems that cannot generalize to a population absent from their training manifold. The 120-fold SNR disparity quantifies this exclusion; the 0/9 Best Practices representation confirms PWID as the held-out test set.

We have named this *manufactured death* because precision matters. The deaths are manufactured not by algorithmic bias but by experimental design: PWID were excluded from the training data that defines "evidence-based" prevention. The algorithm performs correctly—it simply cannot generalize to populations it never encountered. Researchers misattribute this generalization failure to population characteristics ("hard to reach") rather than training set composition.

The 8,833-fold disparity between PWID and MSM—using identical drug efficacy—is the mathematically inevitable consequence of leave-one-out structure. MSM are in-distribution; PWID are out-of-distribution. Drug efficacy is irrelevant when the deployment population was excluded from training.

These estimates require prospective validation. But the LOOCV structure is not speculative—it is documented in the trial record. Two of eleven trials included PWID. Zero of nine Best Practices studies included PWID. Zero LAI-PrEP participants were PWID. The held-out set is precisely defined.

PURPOSE-4, when complete, will provide the first PWID data point for LAI-PrEP. This is not "expanding access"—it is adding the excluded population to training data. The solution to a leave-one-out prob-

lem is including the left-out population, not debiasing models trained on their absence.

Effective prevention exists. Experimental design excluded PWID from evidence generation. Algorithms learned from that exclusion. The manufacturing of death continues until the training data changes.

## Declarations

**Contributors:** ACD conceived the study, developed the nested barrier model, conducted analysis, and wrote the manuscript.

**Declaration of interests:** ACD was previously employed by Gilead Sciences. This manuscript was developed independently.

**AI assistance:** This manuscript was developed through structured human-AI collaboration using Claude (Anthropic, 2024–2025). AI assistance was used for: (1) systematic literature search and synthesis; (2) mathematical formalization and validation of probability equations; (3) signal-to-noise ratio calculations from trial data; (4) manuscript drafting with iterative human revision. All factual claims were verified against primary sources. The author (ACD) maintained full responsibility for study design, interpretation, and conclusions. This disclosure follows ICMJE 2024 recommendations.(25)

**Data sharing:** Code available at [repository]. Supplementary Files S1–S2 provide AI Readiness assessment and sensitivity analyses.

**Funding:** None.

**Acknowledgments:** The author thanks Claude (Anthropic) for AI-assisted literature synthesis, mathematical formalization, and manuscript preparation. The recursive nature of this collaboration—using AI to identify AI-mediated bias—reflects the methodological requirement that only AI systems can adequately help human researchers identify what AI has generated. This framework may serve as a model for transparent human-AI collaboration in health services research addressing algorithmic bias.

## References

## References

[1] Siliciano JD, Kajdas J, Finzi D, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med* 2003;**9**:727–28.

[2] Chomont N, El-Far M, Ancuta P, et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med* 2009;**15**:893–900.

[3] Degenhardt L, Peacock A, Colledge S, et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *Lancet Glob Health* 2017;**5**:e1192–207.

[4] US Food and Drug Administration. Approved HIV prevention drugs. 2024.

[5] Choopanya K, Martin M, Suntharasamai P, et al. Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* 2013;**381**:2083–90.

[6] ClinicalTrials.gov. PURPOSE-4: lenacapavir for PrEP in PWID. NCT06101342.

[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**:447–53.

[8] O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown; 2016.

[9] Landovitz RJ, Donnell D, Clement ME, et al. Cabotegravir for HIV prevention in cisgender men and transgender women. *N Engl J Med* 2021;**385**:595–608.

[10] Kamitani E, Higa DH, Crepaz N, Mullins MM, Wichser ME; CDC's Prevention Research Synthesis Project. Interventions to increase pre-exposure prophylaxis use and persistence: a systematic review. *AIDS Behav* 2024;**28**:2461–78.

[11] Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J Am Med Inform Assoc* 2024;**31**:1172–83.

[12] Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Commun Med* 2022;**2**:149.

[13] Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng* 2024;**8**:719–42.

[14] Butt JH, Nielsen OW, Kristensen SL, et al. Are treatment services ready for the use of big data analytics and AI in managing opioid use disorder? *JMIR Med Inform* 2025;**13**:e52555.

[15] Guerrero EG, Fenwick K, Kong Y, Grella C, D'Aunno T. Using machine learning to advance disparities research: subgroup analyses of access to opioid treatment. *Health Serv Res* 2022;**57**:411–21.

[16] Fletcher RR, Nakeshimana A, Olubeko O. Addressing algorithmic bias and the perpetuation of health inequities. *Lancet Digit Health* 2024;**6**:e870–82.

[17] Mistler CB, Copenhaver MM, Shrestha R. The pre-exposure prophylaxis (PrEP) care cascade in people who inject drugs: a systematic review. *AIDS Behav* 2021;**25**:1490–506.

[18] Biello KB, Bazzi AR, Mimiaga MJ, et al. Perspectives on HIV pre-exposure prophylaxis (PrEP) utilization and related intervention needs among people who inject drugs. *Harm Reduct J* 2018;**15**:55.

[19] Bazzi AR, Biancarelli DL, Childs E, et al. Limited knowledge and mixed interest in pre-exposure prophylaxis for HIV prevention among people who inject drugs. *AIDS Patient Care STDS* 2018;**32**:529–37.

[20] DeBeck K, Cheng T, Montaner JS, et al. HIV and the criminalisation of drug use among people who inject drugs: a systematic review. *Lancet HIV* 2017;**4**:e357–74.

[21] Altice FL, Azbel L, Stone J, et al. The perfect storm: incarceration and the high-risk environment perpetuating transmission of HIV, hepatitis C virus, and tuberculosis in Eastern Europe and Central Asia. *Lancet* 2016;**388**:1228–48.

[22] Strathdee SA, Kuo I, El-Bassel N, et al. Preventing HIV outbreaks among people who inject drugs in the United States: plus ça change, plus c'est la même chose. *AIDS* 2020;**34**:1997–2005.

[23] Baugher AR, Wejnert C, Kanny D, et al. Are we ending the HIV epidemic among persons who inject drugs?: key findings from 19 US cities. *AIDS* 2025;**39**:1813–19.

[24] Demidont AC. AI Readiness in Healthcare: Framework for Computational Validity and Clinical Deployment. *Viruses* 2025; Supplementary File S3.

[25] International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated January 2024.

[26] Cooke SJ, Nguyen VM, Young N, et al. Human-AI collaboration to identify literature for evidence synthesis. *Cell Rep Sustain* 2024;**1**:100132.

[27] Collins KM, Sucholutsky I, Bhatt U, et al. Building machines that learn and think with people. *Nat Hum Behav* 2024;**8**:1851–63.

[28] Gomez C, Cho SM, Ke S, Huang CM, Unberath M. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making. *Front Comput Sci* 2024;**6**:1521066.

[29] Liang W, Zhang Y, Cao H, et al. Monitoring AI-modified content at scale using LLMs. *Nat Hum Behav* 2025. doi:10.1038/s41562-025-02273-8.

[30] Hamed AA, Zachara-Szymanska M, Wu X. Safeguarding authenticity for mitigating the harms of generative AI. *iScience* 2024;**27**:108782.

[31] Pearson AT, Li A, Hosseini FM, et al. Characterizing the increase in AI content detection in oncology scientific abstracts. *JCO Clin Cancer Inform* 2024;**8**:e2400077.

[32] Kobak D, Gonzalez-Marquez R, Horvát EA, Lause J. Delving into ChatGPT usage in academic writing through excess vocabulary. *Sci Adv* 2024. doi:10.1126/sciadv.adn5490.

[33] Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al. Testing of detection tools for AI-generated text. *Int J Educ Integr* 2023;**19**:26.