Drug and Alcohol Review

WILEY | APSAD

ORIGINAL PAPER

From Clinical Trials to Real-World Impact: Introducing a Computational Framework to Detect Endpoint Bias in Opioid Use Disorder Research

Gabriel J. Odom[1] | Laura Brandt[2] | Aaron Marker[3] | Salvatore Giorgi[3] | Ganesh Jainarain[2] | H. Andrew Schwartz[3] | Larry Au[4] | Clinton Castro[5] | The ENDPOINT Consortium

[1]Department of Biostatistics, Florida International University, Miami, Florida, USA | [2]Department of Psychology, The City College of New York, New York, New York, USA | [3]Department of Computer Science, Stony Brook University, Stony Brook, New York, USA | [4]Department of Sociology, The City College of New York, New York, New York, USA | [5]The Information School, The University of Wisconsin—Madison, Madison, Wisconsin, USA

Correspondence: Laura Brandt (lbrandt@ccny.cuny.edu)

Received: 7 March 2025 | **Revised:** 15 November 2025 | **Accepted:** 20 November 2025

Keywords: algorithmic bias | demographic parity | open-source software | opioid use disorder | performance variance

ABSTRACT

Introduction: Clinical trial endpoints are a 'finite sequence of instructions to perform a task' (measure treatment effectiveness), making them algorithms. Consequently, they may exhibit algorithmic bias: internal and external performance can vary across demographic groups, impacting fairness, validity and clinical decision-making.

Methods: We developed the open-source Detecting Algorithmic Bias (DAB) Pipeline in Python to identify endpoint 'performance variance'—a specific algorithmic bias—as the proportion of minority participants changes. This pipeline assesses internal performance (on demographically matched test data) and external performance (on demographically diverse validation data) using metrics including F1 scores and area under the receiver operating characteristic curve (AUROC). We applied it to representative opioid use disorder (OUD) trial endpoints.

Results: F1 scores remained stable across minority representation levels, suggesting consistency in precision-recall balance (F1) despite demographic shifts. Conversely, AUROC measures were more sensitive, revealing significant performance variance. Training on demographically homogeneous populations boosted internal performance (accuracy within similar cohorts) but critically compromised external generalisability (accuracy within diverse cohorts). This pattern reveals an 'endpoint bias trade-off': optimising performance for homogeneous populations vs. having generalisable performance for the real world.

Discussion and Conclusions: Consistently performing endpoints for one demographic profile may lose generalisability during population shifts, potentially introducing endpoint bias. Increasing minority representation in the training data consistently improved generalisability. The endpoint bias trade-off reinforces the importance of diverse recruitment in OUD trials. The DAB Pipeline helps researchers systematically pinpoint when an endpoint may suffer 'performance variance' (i.e., bias). As an open-source tool, it promotes transparent endpoint evaluation and supports selecting demographically invariant OUD endpoints.

Gabriel J. Odom and Laura Brandt contributed equally to this article.

1 | Introduction

1.1 | Clinical Trial Endpoints as Algorithms

Much like in the treatment of other psychiatric and medical conditions, opioid use disorder (OUD) clinical trials rely on specific *endpoints*—measurable outcomes used to gauge the effect of an intervention—when evaluating the safety and efficacy of medications, behavioural therapies and their combinations. However, there is no objective 'gold standard' to evaluate if a patient has recovered from OUD [1–3], and measures of treatment success may include frequency of substance (non)use, abstinence period(s) length, craving for primary substance used and risk of relapse, among others.

A clinical trial endpoint is determined by a finite sequence of instructions for performing a task—and thus is an *algorithm* [4]. For example, Krupitsky et al. [5] measured 'confirmed abstinence during study weeks 5–24'. In OUD clinical trials, 'confirmed abstinence' typically requires one or more urine drug screens (UDS) negative for non-treatment opioids. Regarding this example, the ordered set of instructions to calculate abstinence for a participant at trial completion is: (i) partition all clinic visits into study weeks; (ii) for each study week, mark if the patient (a) supplied an opioid-negative UDS, (b) supplied a positive UDS, or (c) failed to supply a UDS; (iii) for each patient, count the number of weeks with exclusively negative UDS during the 20-week period from study week 5 to study week 24; and (iv) mark the participant as 'abstinent' if they had 20 weeks of exclusively opioid-negative UDS, 'non-abstinent' otherwise. This set of instructions is finite, having four steps. It is a sequence: steps must be done in order. Each step is a set of instructions: some involve informatics, chemistry or mathematics, but they are instructions. This 4-step ordered set of instructions performs a single task: mark if a clinical trial participant achieved 'abstinence' from opioids.

We acknowledge that describing endpoints as algorithms may be unconventional outside computer science, but our use of the term is precise: endpoints are finite, rule-based procedures that transform raw data into clinical outcomes. The purpose of this framing is not to conflate clinical endpoints with computational Artificial Intelligence (AI) algorithms, but to enable the application of established fairness diagnostics to assess whether endpoint definitions systematically disadvantage certain groups. As demonstrated in Odom/Brandt et al. [6], most—if not all—clinical trial endpoints used to evaluate treatments for OUD can be expressed as finite sequences of data transformation and analysis instructions designed to measure some indicator of recovery.

Thus, clinical trial endpoints function as algorithms and, as a result, are susceptible to *algorithmic bias* [7–10], which we define as 'systematic deviation in algorithm output, performance, or impact, relative to a contextually salient standard' [11, 12]. When algorithms deviate from such a standard, their performance varies across demographic groups in ways that may affect fairness, validity, and subsequent decision-making. In the context of clinical trials, we define this as *endpoint bias*—a form of algorithmic bias that emerges when the selection, definition or operationalisation of trial endpoints systematically favours certain outcomes, populations or interpretations over others.

1.2 | Our Perspective on Bias in Clinical Trials: The Case of Performance Variance

We focus on a specific form of endpoint bias, *performance variance*, which we define as 'a change [variance] in model performance as the proportion of the samples shifts between the minority and majority classes'. We take *model performance* to mean measurable characteristics of the model, specifically metrics of error. The minority and majority classes are defined by a demographic attribute of interest. Clinical trial demographics are measured across many attributes, such as race, ethnicity, age, sex (assigned at birth), income strata, education level and others. Some of these are *protected attributes* [13], related to their status in laws against discrimination.

We consider two definitions of 'performance':

1. The ability of a model to predict data that resembles its training data in terms of demographic composition, which we refer to as *internal performance (in)variance*. This corresponds to questions such as 'If an endpoint was validated in a past clinical cohort, will it yield good model performance when applied to a recently recruited cohort with similar demographics?'
2. The ability of a model to predict data with a different demographic composition, which we refer to as *external performance (in)variance*. This addresses, for example, whether an endpoint developed in a predominantly non-Hispanic white cohort will generalise to a more diverse population.

Here, bias does not imply intentional discrimination but rather systematic differences in how endpoints perform across different groups [14]. Just as predictive models can exhibit disparities in accuracy depending on the demographics of the data they are trained on [15–17], endpoint definitions can reflect underlying biases in how treatment success is measured. For example, an endpoint requiring prolonged abstinence based on UDS results may disadvantage participants with fewer resources for frequent clinic visits [18–20]. These biases can have real consequences. If an endpoint disproportionately classifies one group as 'non-responders' due to structural factors rather than true differences in treatment effect, it can lead to misleading conclusions about efficacy and reinforce disparities in care. Recognising endpoints as algorithms allows us to systematically examine whether definitions of treatment success introduce disparities.

1.3 | The Detecting Algorithmic Bias Pipeline

The Detecting Algorithmic Bias (DAB) pipeline is an open-source software tool (available in Python 3.12 at https://github.com/CTN-0094/Pipeline) designed to detect whether clinical trial endpoints for OUD exhibit performance variance—in other words, whether an endpoint's performance metrics vary as we shift the demographic makeup of a sample. Because these endpoints define treatment success (e.g., 'abstinent' vs.

'non-abstinent'), each step in the DAB pipeline tracks how well a given endpoint generalises across different protected attributes (e.g., race/ethnicity, sex or age). Conceptually, this pipeline works as an 'assembly line' of discrete steps, transforming raw data into evidence about endpoint bias (see 'Methods' for details)—to see if the chosen endpoint's predictive accuracy holds steady (performance invariance) or falters (performance variance).

In practical terms, the DAB pipeline is suited to smaller, fragmented datasets common in clinical research. We acknowledge that in many AI/machine-learning (ML) contexts, models have access to learn from tens of thousands of samples or more from large hospital or census data sets, and the design of a pipeline to detect algorithmic bias in that context would be different. However, because the motivation of this work was to help clinical trialists choose treatment outcomes based on empirical properties (rather than historic prevalence, reason or even whimsy), sample size limitations are built intrinsically into this pipeline. Rather than relying on large validation sets, the pipeline is built on techniques in bootstrapping and cross-validation. The result is a 'portfolio' of performance metrics—both for internally matched participants (similar demographics to the training data) and an external validation set (different demographic makeup).

The DAB pipeline is provided as an open-source, reproducible workflow that enables researchers to systematically evaluate whether endpoint definitions exhibit performance variance across demographic groups, thereby offering an empirical approach to detecting potential bias in treatment outcome measurement. Although we demonstrate the DAB Pipeline using clinical trial endpoints, the framework was designed as a meta-AI method to evaluate fairness and generalisability in AI/ML applications more broadly. In this way, it provides a diagnostic tool for detecting performance degradation in imbalanced or heterogeneous datasets and complements ongoing efforts to build AI-enabled decision support systems in substance use and addiction research.

1.4 | Study Aims

The goal of this study is to examine endpoint bias in OUD clinical trials and introduce a framework for assessing its impact on treatment evaluation. Specifically, we aim to investigate performance variance as a key mechanism of endpoint bias, focusing on how endpoint performance differs across demographic groups. In this application we focus on race/ethnicity as the protected attribute of interest. We stress that this use does not imply a causal interpretation; rather, race/ethnicity functions as a grouping variable to test whether endpoint performance varies across demographic strata. Using synthetic cohort simulations, we demonstrate how the DAB pipeline can quantify internal performance variance (performance on demographically similar cohorts) and external performance variance (generalisability to more diverse populations). By recognising the algorithmic nature of endpoints and applying quantitative bias assessments, this study seeks to improve the effectiveness of treatment outcome measurement in OUD clinical research.

We provide Supporting Information detailing the development process of the DAB pipeline for interested readers seeking methodological insights and practical guidance: we discuss our guiding philosophy for solution development (Section 1); we contrast our approach—designed for clinical trials with limited sample sizes—with a method typically used for larger datasets (Section 2); we outline how the pipeline generates synthetic cohorts (Section 3); and we demonstrate the pipeline's practical utility by applying fairness diagnostics to two case studies of commonly used endpoints (Section 4), highlighting its immediate applicability for clinical researchers.

2 | Methods

The DAB pipeline consists of three main components: inputs, model evaluation, and outputs. Each component plays a critical role in assessing performance variance in clinical trial endpoints. Details on our previous work that informed the development of the DAB pipeline, as well as our guiding philosophy for solution development, are provided in Section 1 of the Supporting Information. For additional clarity, we contrast our approach, detailed below, with a competing method, the 'virtual twin' approach, in Section 2 of the Supporting Information. Figure 1 provides a visual overview of the DAB pipeline, illustrating the process from data sampling to evaluation.

2.1 | The DAB Pipeline: Inputs

The DAB Pipeline takes three inputs:

1. A set of treatment outcomes;
2. A rich data set including binary protected attributes with sufficient and independent features (including demographics, clinical information, behavioural assessments, genetics measurements or other informative features or important controls/confounding effects); and
3. A conceptual statistical, ML or AI model to predict the outcomes from the independent features.

For example, the treatment outcomes: (1) could be the number of study weeks until two consecutive clinic visits have been missed (a metric for relapse) over a data set, (2) of male and female participants of a specific clinical trial (e.g., CTN-0027 of Saxon et al. [21]) using a predictive model that is, (3) a negative binomial regression with stepwise feature selection.

2.1.1 | Current Default Values of the Pipeline

We use an existing clinical trials data set with sufficient samples across each of the protected attributes of interest, the harmonised CTN-0094 data set (see Balise et al. [22] for metadata). This is the largest public OUD clinical trial data set and includes clinical cohorts from three clinical trials that tested the efficacy of various medication-assisted treatments. It has 3560 samples across 23 tables in a fifth normal form database and seven Supporting Information tables of engineered variables. This database is harmonised, de-identified, open source and freely

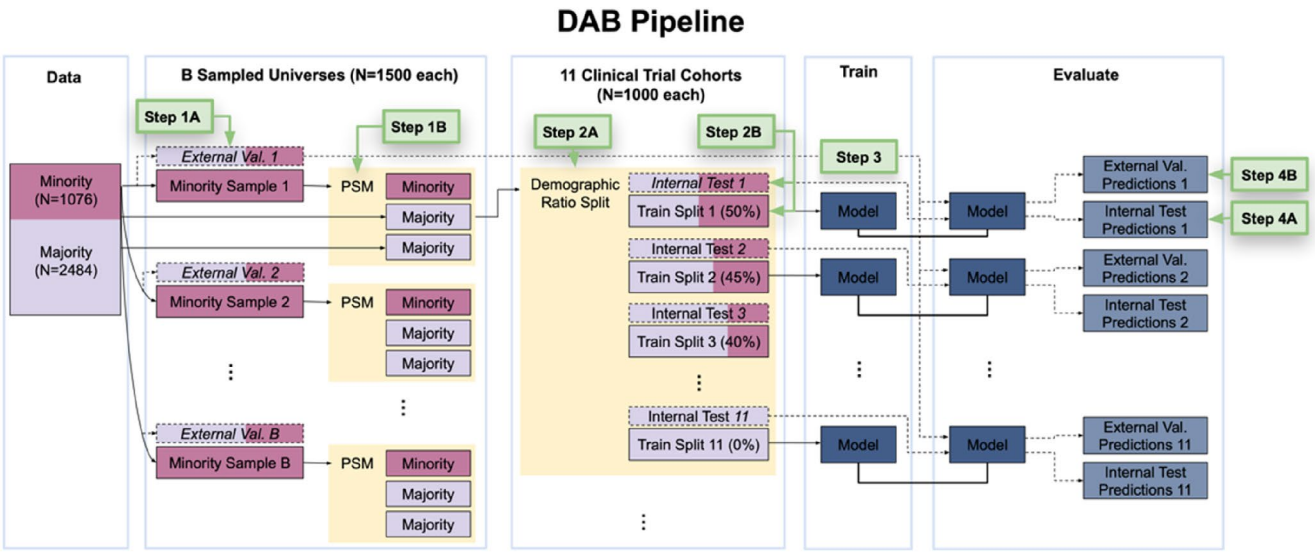

FIGURE 1 | A visual overview of the detecting algorithmic bias (DAB) pipeline, illustrating the process from data sampling to evaluation. This figure shows the 'flow' of the data, from creating 'universes' by random sampling to the modelling results for those universes. These model results enable visualisation of performance variance of clinical trial endpoints due to changes in demographic composition within trial cohorts.

available in two data packages for the R programming language [23, 24]. We focus on racial/ethnic minority status as the protected attribute of interest, so we note that there are 2484 non-Hispanic white participants and 1076 participants belonging to a racial/ethnic minority group in the CTN-0094 database. That said, the DAB pipeline will work for any binary protected attribute, and we are expanding it to account for continuous and categorical protected attributes with three or more groups.

As discussed in Brandt/Odom et al. [18], there have been over 50 subject-specific clinical trial treatment endpoints published since the 1970s, but there is no consensus on which endpoints are 'the best' (or what that would mean). While this pipeline supports the use of any subject-specific treatment endpoint, the default choices include a representative sample of the endpoints included in our CTNote library [6]. Specifically, we found that the main characteristics of treatment endpoint algorithms can be divided into two axes: the traditional groupings of abstinence, relapse and use reduction, and also a novel dimension which considers how many (few [1, 2] or many [3+]) UDS values are used to make a clinical determination. We then overlaid binary treatment endpoints onto a Latin Square lattice of these two axes and chose representative endpoints for each cell in the square. These endpoints are shown in Table 1.

2.1.2 | Choice of Model and Model Performance Metrics

Once we have selected the treatment outcome, protected attribute and predictors, and conceptual model, we need to choose numeric model performance metrics. For example, if the treatment outcome was a continuously valued protein biomarker related to substance use [25] and the conceptual model was multiple least squares regression, then R Squared is the standard metric to measure model accuracy. However, if the treatment outcome was a binary indicator of end of treatment abstinence and the conceptual model was logistic regression or a neural network, then area under the receiver operating characteristics curve (AUROC) or *F1* score would be common metrics to measure model accuracy. Under most circumstances, the metric should only be compared as the protected attribute changes, but all other inputs (treatment outcome, predictors and conceptual model) are held constant.

2.2 | The DAB Pipeline: Process

At this point, we have a treatment outcome of interest: a binary protected attribute and other independent predictors, covariates and recorded confounding variables; an appropriate conceptual predictive/regression model; and a metric to evaluate said model's ability to predict the treatment outcome. We now give the general process of how to assess the performance variance of the treatment outcome to the protected attribute (details in Supporting Information Section 3). For a visual overview, see the DAB pipeline flowchart in Figure 1.

Step 1. (**A**) We start by drawing at random m samples to hold aside as an external validation set. (**B**) Then, we randomly draw n samples of the 'minority' category (based on the protected attribute of interest) from the clinical trial data set, and we draw $2n$ propensity-score matched samples of the 'majority' category [26]. To measure *internal performance variance* for similar data, we then pretend that these $n+2n$ samples represent all the participants available for recruitment to clinical trials in this universe. To measure *external performance variance* for different data, the models trained on subsets of these $n+2n$ samples will be used to predict the m validation samples.

Step 2 (**A**) For both 'performance' definitions ('ability to predict similar data' and 'ability to predict new data'), we begin by 'recruiting' clinical trial participants from the $n+2n$ people in our universe. The first synthetic clinical trial (cohort 1 for the sample drawn in **Step 1**) will be all n minority persons and half of the two sets of n majority persons; this will yield a clinical trial cohort of size $2n$ with demographic parity. Using this 50–50 minority-majority cohort, (**B**) we create train-test data splits,

TABLE 1 | Representative opioid use disorder treatment endpoints by endpoint category and number of urine drug screens (UDS) used to determine 'treatment success'.

No. of UDS used to make determination	Category								
	Abstinence			Relapse			Use reduction		
	Endpoint definition	Used by	CTNote URL	Endpoint definition	Used by	CTNote URL	Endpoint definition	Used by	CTNote URL
1–2 UDS	Confirmed opioid abstinence during weeks 5–24 based on UDS	Krupitsky et al. [5]	https://ctn-0094.github.io/CTNote/articles/library_abstinence_20220711.html#krupitsky-et-al--2011-a	2 consecutive positive UDS following 4 weeks of treatment; missing is positive	Johnson et al. [30]	https://ctn-0094.github.io/CTNote/articles/library_relapse_20220711.html#johnson_controlled_1992	N/A[a]	N/A[a]	N/A[a]
3+ UDS	13 consecutive negative UDS (1 month); urine was screened 3 times per week	Ling et al. [31]	https://ctn-0094.github.io/CTNote/articles/library_abstinence_20220711.html#ling_buprenorphine_1998	3 consecutive positive UDS	Krupitsky et al. [32]	https://ctn-0094.github.io/CTNote/articles/library_relapse_20220711.html#krupitsky_naltrexone_2004-and-krupitsky_naltrexone_2006	≥70% negative UDS during the 24-week trial period; missing UDS are excluded	Kosten et al. [33]	https://ctn-0094.github.io/CTNote/articles/library_reduction_20220630.html#kosten_buprenorphine_1993

[a]Clinical trials do not typically report any participant-specific, binary, use-reduction-focused treatment endpoints which are calculated only using a few UDS. Nearly all substance use reduction endpoints look at average use patterns over weeks or months.

(**Step 3**) build the specified prediction model, and (**Step 4**) measure its accuracy by the specified metric in both (**A**) the testing data and (**B**) the external *m* validation samples.

To assess the impact of changing the demographics of the cohort while minimising sources of additional variance, we perturb the demographic composition of the protected attribute *ceteris paribus*. Specifically, for the second synthetic clinical trial, we 'change the recruitment procedure' of the 'cohort', replacing 5% of the minority persons with their propensity-score matched counterparts from the remaining *n* majority samples. With this new 45–55 minority-majority cohort (cohort 2 for the sample drawn in **Step 1**), we repeat **Steps 2B—4**: we re-split the data, rebuild the prediction model, and re-measure the model fit performance in the internal test and external validation data sets. We repeat this process (for the third, fourth, fifth, etc. synthetic clinical trials) to create cohorts with fewer and fewer minority participants.

2.3 | The DAB Pipeline: Outputs

As we slowly shift the clinical trial recruitment demographics from complete 50–50 minority-majority parity to 0–100 majority dominance (**Step 2**) while holding all else constant, we record the performance of the model built in **Step 3** to predict the endpoint of interest for the internal test data (**Step 4A**) and external validation samples (**Step 4B**). If an endpoint maintains near-constant model performance in the *internal test data* as the demographic profile of the training and test data changes (for the chosen protected attribute), then the treatment outcome is *internally performance invariant* to the protected attribute (or *internally performance variant* if the model accuracy changes). If a treatment outcome maintains near-constant model accuracy in the *external validation data* as the demographic composition of the training and test data changes (for the chosen protected attribute), then the treatment outcome is *externally performance invariant* to the protected attribute (or *externally performance variant* if the model accuracy changes).

3 | Results

The DAB pipeline creates tables of model performance statistics across cohort prevalence levels for the chosen protected attribute. Figures 2 and 3 visualise these effects for the representative treatment endpoints introduced in Table 1. Figure 2 displays F1 scores, which are the harmonic mean of precision and recall. We observe that this metric remains relatively constant even as the percentage of minority samples shifts from 0% to nearly 50%.

Figure 3 focuses on AUROC, a measure of a model's overall discriminative ability (ranging from 0.5 to 1, where higher is better). In the Internal panels, AUROC rises as the cohort becomes more demographically uniform, indicating stronger performance on data similar to the training set. However, in the External panels—where the validation set's demographics remain fixed—AUROC declines under more homogeneous training data, suggesting poorer generalisability.

Our analysis reveals a clear pattern: F1 scores remain largely unchanged across demographic compositions, whereas AUROC

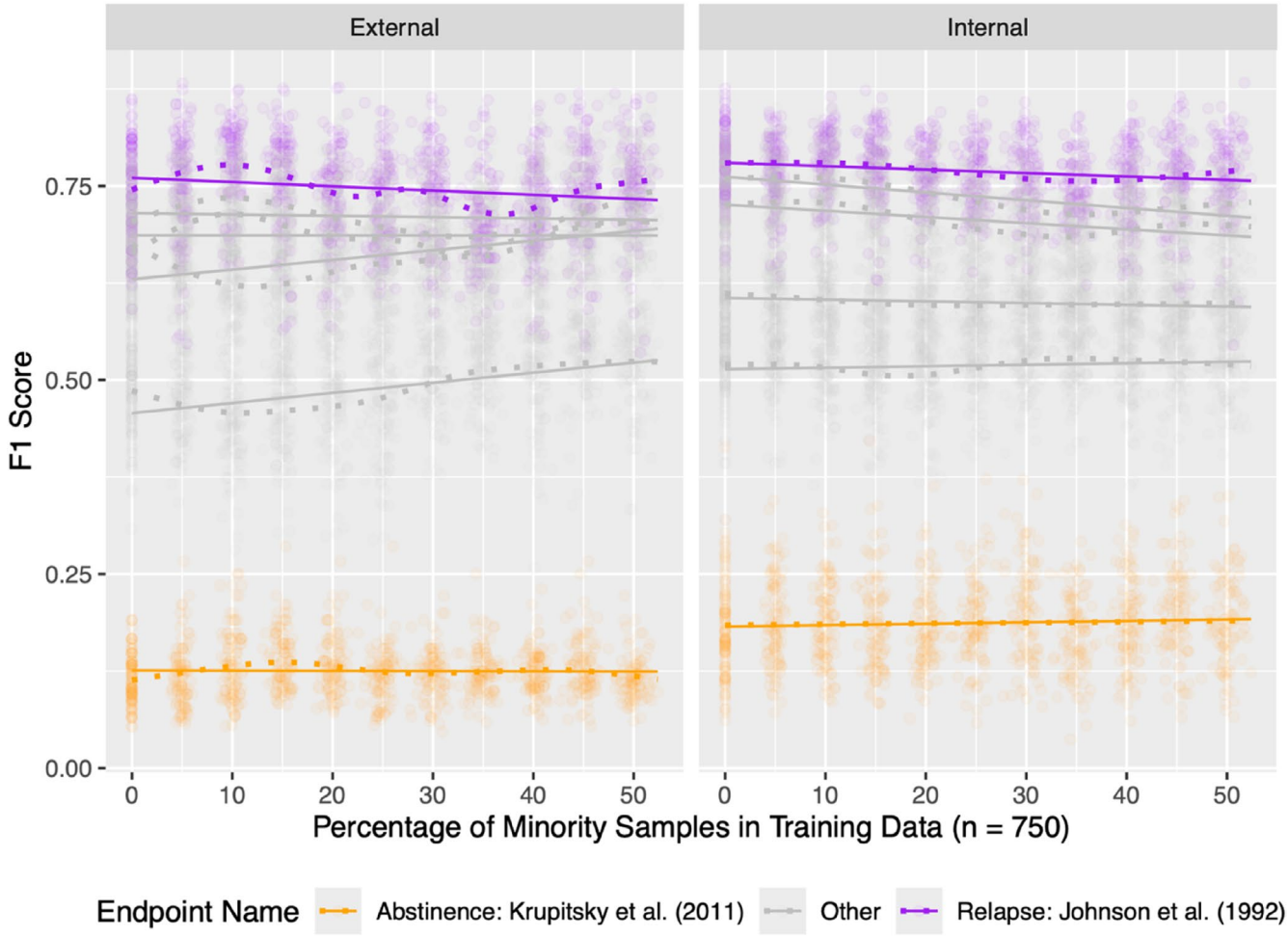

FIGURE 2 | F1 score stability across changing demographic compositions in opioid use disorder endpoints. This figure illustrates how changes in the proportion of minority participants in the training data (*x*-axis) affect F1 scores (*y*-axis, range 0–1) for the endpoints listed in Table 1. The left panel ('External') evaluates a validation set with a fixed 42% minority composition, reflecting approximate U.S. demographics in 2025, while the right panel ('Internal') uses a matched test set whose demographics shift alongside the training data. Each open dot represents a randomly sampled 'universe', with solid (linear) and dashed (LOESS) lines showing best-fit trends. Colours distinguish two use cases: A strict abstinence endpoint (orange) and a two-week relapse endpoint (purple; see Supporting Information Section 4 for more details). Despite varying minority proportions (0% to ~50%), F1 scores remain relatively stable across all endpoints.

shows a dependence on the proportion of minority participants in the training data. Put differently, homogeneous (predominantly non-Hispanic white) training sets lead to higher internal performance but reduce external performance, whereas diverse training cohorts yield better external validation results. Across endpoint definitions, the trends are consistent: a trade-off emerges between maximising model accuracy in internally matched test data and extending that accuracy to external populations with different demographics. These findings underscore the importance of recruiting diverse clinical trial cohorts to bolster real-world applicability—even if it may reduce performance within the trial's own sample. Additional use case results are presented in the Supporting Information (Section 4).

4 | Discussion

Building on prior work to standardise and compare endpoint definitions [6, 18, 27], the DAB pipeline is a systematic method to evaluate whether endpoints exhibit internal or external performance variance based on protected attributes such as race/ethnicity, age and sex. By applying the DAB pipeline to a range of UDS-based endpoints, we found that some commonly used measures exhibited performance variance, meaning their classification of treatment success changed as the demographic composition of the sample varied. Importantly, the performance of these endpoints differed when evaluated under internal versus external performance variance conditions—meaning that an endpoint might appear unbiased within a given clinical trial cohort but fail to generalise to broader populations.

It is important to note that detecting performance variance does not mean an endpoint is inherently bad or invalid. Endpoints may still be clinically useful, especially if they are well aligned with a particular research question or population. For example, an endpoint that exhibits greater performance variance across groups might nonetheless provide valuable diagnostic insight within a specific subgroup, even if its generalisability is limited. The purpose of the DAB pipeline is not to rank endpoints as universally good or bad, but

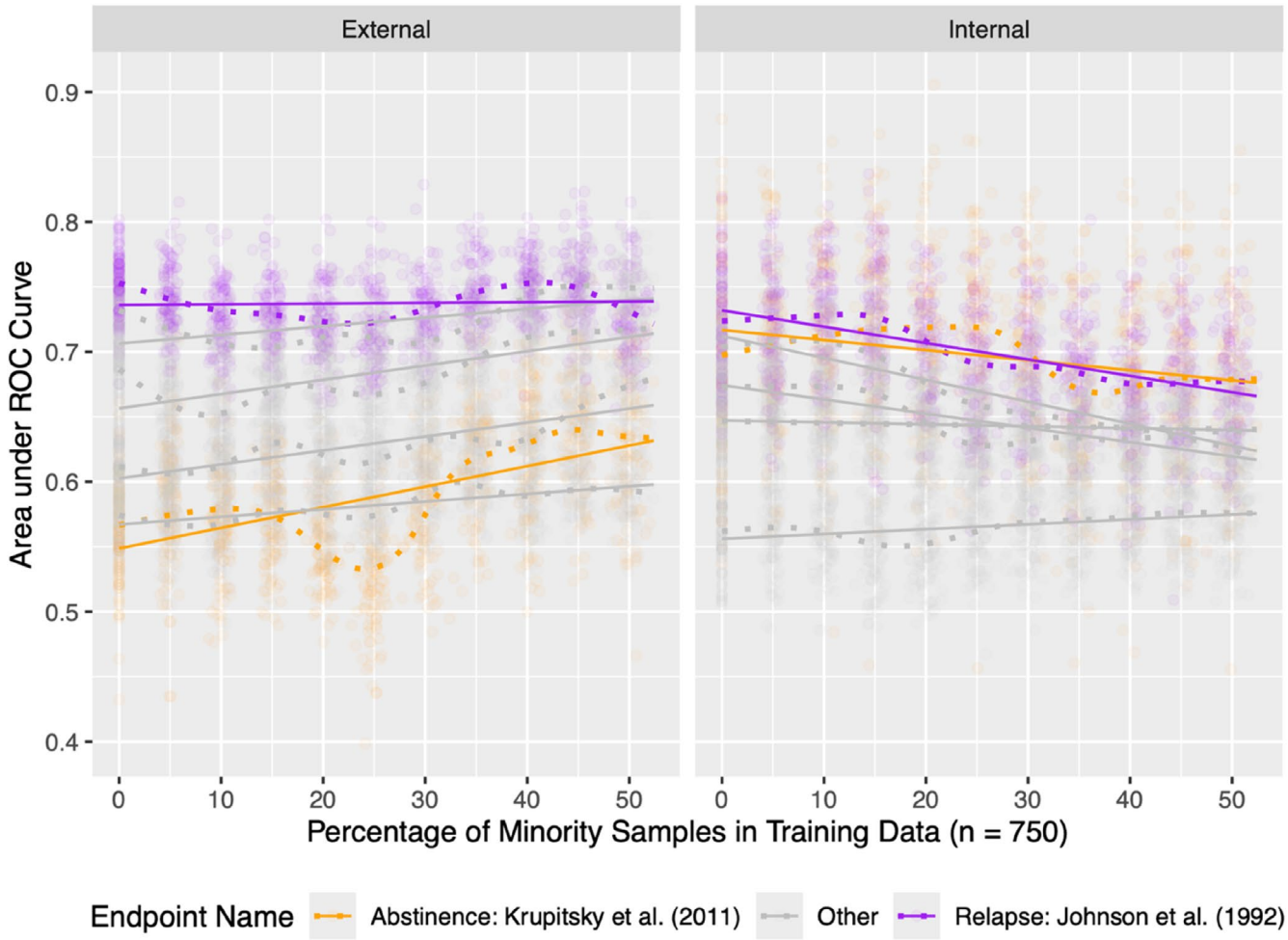

FIGURE 3 | Area under the receiver operating characteristics curve sensitivity to varying minority representation in opioid use disorder endpoints. This figure shows how increasing minority composition in the training data (*x*-axis) impacts Area Under the Receiver Operating Characteristic (AUROC; *y*-axis, range 0.5–1), a measure of a model's overall discriminative ability. The left ('External') panel again fixes the validation set at 42% minority, while the right ('Internal') panel mirrors the training data's changing demographics. Colours distinguish two use cases: A strict abstinence endpoint (orange) and a two-week relapse endpoint (purple; see Supporting Information Section 4 for more details). As minority representation increases in the training data, internal performance (right panel) typically declines, whereas external performance (left panel) improves—highlighting a trade-off between optimising results within a homogeneous cohort and maximising generalisability to more diverse populations.

to make explicit *where* and *for whom* an endpoint's performance changes. Documenting these differences is essential for ensuring fairness, transparency and appropriate endpoint selection in trial design, even when endpoints remain valid for narrower applications.

4.1 | Diverse Cohorts Needed

This finding underscores a fundamental trade-off between clinical trial design and generalisability. Historically, trials prioritised internal validity by recruiting homogeneous cohorts—often young white males—to minimise variability and maximise statistical power for detecting treatment effects [28]. While highly homogeneous cohorts may optimise endpoint reliability within a trial, they risk producing biased or misleading conclusions when applied to real-world, heterogeneous populations (see Figure 3). Conversely, recruiting a more diverse cohort introduces greater variability, reducing power to detect small treatment effects but enhancing the study's ability to predict treatment outcomes in real-world populations. Our findings support this shift in clinical trial philosophy [29]: as the proportion of minority participants in the training data increases, AUROC for external validation data improves, reinforcing the need for diverse recruitment to enhance generalisability. The DAB pipeline provides a systematic framework for evaluating how endpoints behave under these different conditions, offering research teams a tool for endpoint selection.

4.2 | Proposed Use Cases

The DAB pipeline's modular design allows researchers to vary specific input components—*1*. treatment outcome, *2*. predictors, *3*. protected attribute and *4*. conceptual model and performance metrics—while holding others constant. In the near term, we recommend systematically varying one input group at a time to evaluate different sources of performance variance. The pipeline excels at detecting bias due to overfitting, leveraging multiple testing splits, validation sets and repeated

random data resampling to rigorously evaluate endpoints and conceptual models.

A primary use of the pipeline is assessing new and existing treatment outcome algorithms for their appropriateness in OUD clinical trials. Our repository includes commonly used endpoints (e.g., abstinence, use reduction and relapse) measured across binary, count and continuous scales. However, these are not exhaustive, and we recommend testing new endpoints for performance variance before they are incorporated into trial protocols. Retrospective analyses of endpoints from past studies—many of which were chosen based on expert consensus rather than empirical validation—can also benefit from this approach. By varying the outcome definition while holding the data, conceptual models and performance metrics constant, researchers can systematically evaluate endpoint bias.

Another key application is examining whether endpoints validated in OUD trials generalise to other substance use disorders such as cocaine, methamphetamine, cannabinoids, alcohol, or tobacco. Many OUD endpoints rely on UDS, but UDS may not be universally applicable across substance use disorders. By holding endpoints, models and metrics constant while changing the data set to a clinical trial for another substance use disorder with rich UDS data, researchers can assess which endpoint characteristics are UDS-dependent or substance-specific and which are generalisable. Similarly, to explore performance variance across protected attributes beyond race/ethnicity, the pipeline can be applied to clinical trials with rich demographic and socioeconomic data.

Additionally, the pipeline has applications in Artificial Intelligence (AI)/ML research by allowing for systematic variation of the conceptual model and performance metrics. Its modular Python 3.12 architecture separates key steps—data pre-processing, model training/tuning and performance evaluation—enabling researchers to isolate the effects of each. AI/ML researchers can *1*. assess how different data pre-processing strategies affect model performance while keeping the model and metrics fixed, *2*. compare various AI/ML or statistical models by holding pre-processing and evaluation methods constant, *3*. explore the impact of different performance metrics, including accuracy, sensitivity, AUROC, regression R^2, pseudo-R^2 and novel metrics yet to be developed. This last application is particularly relevant given trade-offs between model performance and endpoint performance variance across protected attributes. The pipeline enables us to investigate fundamental AI/ML questions in substance use disorder research, such as how best to quantify and mitigate bias in predictive models.

Beyond evaluating existing endpoints, the DAB pipeline may also support the development of novel or composite endpoints. By systematically identifying where traditional endpoints exhibit performance variance, researchers can combine insights across multiple measures to design endpoints that better capture treatment effects across diverse populations. While clinical trials still require a single primary endpoint, such an empirical process could help trialists select more contextually appropriate outcomes and motivate innovation in defining comprehensive, fair and generalisable endpoints.

Although our demonstration centres on clinical trial endpoints, the framework is agnostic to the underlying machine-learning model and can extend to contexts such as large language models. In such applications, the endpoints correspond to predicted next-word token IDs, the features are the vector representations of preceding text, and demographic attributes could be drawn from the author or the patient being described. The pipeline could then be used to test whether predictions vary systematically by demographic group, which is particularly relevant as zero-shot predictions from large language models are increasingly proposed for use in clinical and psychological research. This example illustrates the flexibility of the DAB framework beyond clinical trial endpoints, underscoring its broader potential for evaluating fairness in AI/ML applications.

4.3 | Limitations

The DAB pipeline is a necessary litmus test of endpoint bias in clinical trial contexts, but not a sufficient test by itself. It is important to note that while the DAB pipeline highlights *where* endpoint performance differs across demographic groups, it does not identify the *causes* of those differences. Attributes such as race or ethnicity are not causal in themselves but may serve as markers for structural, socioeconomic, or health-related factors that influence endpoint performance. Identifying these underlying mechanisms requires complementary methods, such as causal inference approaches or qualitative investigation, which are beyond the scope of the present work.

Moreover, the DAB pipeline is limited by the quality of the data it receives and cannot detect biases introduced during data collection. However, when used appropriately, it can identify bias related to feature selection—though this requires thoughtful preprocessing. Notably, while the absence of detected feature selection bias does not confirm its absence, interdisciplinary teams with diverse perspectives are best equipped to recognise and mitigate such biases. The DAB pipeline provides one diagnostic 'pillar' by quantifying performance variance in endpoints, but it cannot detect biases that originate upstream in trial design, recruitment or data collection. Thus, bias detection must be complemented by careful, interdisciplinary evaluation to contextualise its impact on clinical decision-making, regulatory approval and real-world treatment effectiveness. Human judgement and stakeholder perspectives remain essential to ensure that endpoints are both empirically robust and contextually fair.

5 | Conclusion

The DAB pipeline is a systematic, empirical approach to identifying performance variance in OUD treatment endpoints. As an open-source software tool, it provides researchers with a structured framework to assess endpoint performance across demographic groups, revealing potential biases in how treatment success is measured. The novelty of this work lies not in showing that trial endpoints can differ across populations—a point already acknowledged in the field—but in offering a transparent, reproducible framework for diagnosing such variance systematically. This contribution operationalises longstanding calls for fairness and generalisability in trial design by providing a concrete tool to evaluate these principles in practice. Our findings demonstrate that endpoint definitions vary in their

generalisability, with some performing well within a given clinical trial but failing to replicate across broader, more diverse populations. By making the DAB pipeline publicly available, we aim to foster transparency and accountability in clinical research. While the tool cannot replace expert judgement it provides an empirical foundation for evaluating endpoint fairness and guiding discussions on improving trial methodologies. Ensuring that treatment endpoints are both valid and representative requires continuous scrutiny, empirical testing and engagement with diverse stakeholders, ultimately advancing the measurement of treatment outcomes in OUD and beyond.

Author Contributions

Gabriel J. Odom and Laura Brandt jointly conceptualized the study and drafted the manuscript. Gabriel J. Odom conducted all statistical analyses. Ganesh Jainarain developed the code under the supervision of Gabriel J. Odom, Laura Brandt, and Salvatore Giorgi. Aaron Marker, Salvatore Giorgi, H. Andrew Schwartz, Larry Au, and Clinton Castro contributed to the development of the study approach, provided critical intellectual input, and reviewed and edited the manuscript. All authors approved the final version.

Acknowledgements

We would like to express our sincere gratitude to the members of our **EN**dpoint Science for **D**ata-Driven **P**recision in **O**pioid Use **I**ntervention and **T**rials (ENDPOINT) Consortium, whose critical contributions have shaped our overarching vision: to advance the standardisation, refinement and transparency of empirically guided endpoint selection in substance use disorder (SUD) clinical trials. Their insights were instrumental in conceptualising the Detecting Algorithmic Bias (DAB) pipeline and its role in improving fairness and validity in treatment outcome measurement. In addition to the authors of this paper, the ENDPOINT Consortium includes Sean X. Luo, MD (Columbia University, Department of Psychiatry), Raymond R. Balise, Ph.D. (University of Miami, Department of Public Health Sciences, Division of Biostatistics and Bioinformatics), Daniel J. Feaster, Ph.D. (University of Miami, Department of Public Health Sciences, Division of Biostatistics and Bioinformatics), and Suky Martinez (Johns Hopkins University, Behavioural Pharmacology Research Unit). We also recognise the invaluable guidance of our advisory board, whose six members played a key role in shaping the pipeline's development and our broader research agenda. Over the course of 2 years, they provided critical feedback during four dedicated meetings. Their collective expertise spans multiple domains, including clinical psychology, behavioural health, Indigenous perspectives, peer recovery, social work and community outreach. Their diverse insights have enriched our work, ensuring its relevance and applicability to both research and real-world treatment settings. Given that several members have lived experience with opioid use disorder, we respect their request for confidentiality. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. This research was also funded by NIMHD FIU-RCMI Pilot AWD000000009108.

Funding

This work was supported by National Institutes of Health, 1OT2OD032581-01, NIMHD FIU-RCMI AWD000000009108.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Comprehensive R Archive Network at https://cran.r-project.org/package=public.ctn0094data.

References

1. B. E. Biondi, X. Zheng, C. A. Frank, I. Petrakis, and S. A. Springer, "A Literature Review Examining Primary Outcomes of Medication Treatment Studies for Opioid Use Disorder: What Outcome Should be Used to Measure Opioid Treatment Success?," *American Journal on Addictions* 29, no. 4 (2020): 249–267.

2. B. B. Dennis, N. Sanger, M. Bawor, et al., "A Call for Consensus in Defining Efficacy in Clinical Trials for Opioid Addiction: Combined Results From a Systematic Review and Qualitative Study in Patients Receiving Pharmacological Assisted Therapy for Opioid Use Disorder," *Trials* 21, no. 1 (2020): 30.

3. A. B. Laudet, "What Does Recovery Mean to You? Lessons From the Recovery Experience for Research and Practice," *Journal of Substance Abuse Treatment* 33, no. 3 (2007): 243–256.

4. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. (MIT Press, 2009), 1312.

5. E. Krupitsky, E. V. Nunes, W. Ling, A. Illeperuma, D. R. Gastfriend, and B. L. Silverman, "Injectable Extended-Release Naltrexone for Opioid Dependence: A Double-Blind, Placebo-Controlled, Multicentre Randomised Trial," *Lancet* 377, no. 9776 (2011): 1506–1513.

6. G. J. Odom, L. Brandt, C. Castro, et al., "Capturing Drug Use Patterns at a Glance: An n-Ary Word Sufficient Statistic for Repeated Univariate Categorical Values," *PLoS One* 18, no. 9 (2023): e0291248.

7. K. Lum, Y. Zhang, and A. Bower, "De-biasing 'bias' measurement," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), 379–389.

8. U. Gal, T. B. Jensen, and M. K. Stein, "People Analytics in the Age of Big Data: An Agenda for IS Research," in *ICIS 2017 Proceedings* (Association for Information Systems. AIS Electronic Library (AISeL), 2017).

9. I. Asadi Someh, M. Davern, C. Breidbach, and G. Shanks, "Ethical Issues in Big Data Analytics: A Stakeholder Perspective," *Communications of the Association for Information Systems* 44 (2019): 718–747.

10. N. T. Lee, "Detecting Racial Bias in Algorithms and Machine Learning," *Journal of Information, Communication and Ethics in Society* 16, no. 3 (2018): 252–260.

11. S. Fazelpour and D. Danks, "Algorithmic Bias: Senses, Sources, Solutions," *Philosophy Compass* 16, no. 8 (2021): e12760.

12. T. Kelly, *Bias: A Philosophical Study* (Oxford University Press, 2023), 288.

13. Wex Definitions Project, "Protected Characteristic," in *The Legal Information Institute* (Cornell Law School, 2020).

14. E. Petersen, S. Holm, M. Ganz, and A. Feragen, "The Path Toward Equal Performance in Medical Machine Learning," *Patterns* 4, no. 7 (2023): 100790.

15. C. L. Andaur Navarro, J. A. A. Damen, T. Takada, et al., "Risk of Bias in Studies on Prediction Models Developed Using Supervised Machine Learning Techniques: Systematic Review," *BMJ* 20 (2021): n2281.

16. R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review," *JAMA Dermatology* 157, no. 11 (2021): 1362–1369.

17. T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A Deeper Look at Dataset Bias," in *Domain Adaptation in Computer Vision*

Applications, ed. G. Csurka (Springer International Publishing, 2017), 37–55.

18. L. Brandt, G. J. Odom, M. Hu, C. Castro, R. R. Balise, and the CTN-0094 Team, "Empirically Contrasting Urine Drug Screening-Based Opioid Use Disorder Treatment Outcome Definitions," *Addiction* 119, no. 7 (2024): 1289–1300.

19. K. A. Morin, J. R. Dabous, F. Vojtesek, and D. Marsh, "Evaluating the Association Between Urine Drug Screening Frequency and Retention in Opioid Agonist Treatment in Ontario, Canada: A Retrospective Cohort Study," *BMJ Open* 12, no. 10 (2022): e060857.

20. A. J. Saxon, "Short-Acting, Full Agonist Opioids During Initiation of Opioid Agonist Treatment in the Fentanyl Era," *JAMA Network Open* 7, no. 5 (2024): e2411398.

21. A. J. Saxon, W. Ling, M. Hillhouse, et al., "Buprenorphine/Naloxone and Methadone Effects on Laboratory Indices of Liver Health: A Randomized Trial," *Drug and Alcohol Dependence* 128, no. 1–2 (2013): 71–76.

22. R. R. Balise, M. C. Hu, A. R. Calderon, et al., "Data Cleaning and Harmonization of Clinical Trial Data: Medication-Assisted Treatment for Opioid Use Disorder," *PLoS One* 19, no. 11 (2024): e0312695.

23. R. R. Balise, G. J. Odom, and A. Calderon, "public.ctn0094data: De-Identified Data from CTN-0094," 2023, https://CRAN.R-project.org/package=public.ctn0094data.

24. G. J. Odom and R. R. Balise, "public.ctn0094extra: Helper Files for the CTN-0094 Relational Database," 2023, https://CRAN.R-project.org/package=public.ctn0094extra.

25. L. Wang, N. Wu, T. Y. Zhao, and J. Li, "The Potential Biomarkers of Drug Addiction: Proteomic and Metabolomics Challenges," *Biomarkers* 21, no. 8 (2016): 678–685.

26. Q. Y. Zhao, J. C. Luo, Y. Su, Y. J. Zhang, G. W. Tu, and Z. Luo, "Propensity Score Matching With R: Conventional Methods and New Features," *Annals of Translational Medicine* 9, no. 9 (2021): 812.

27. G. J. Odom, L. Brandt, R. R. Balise, and L. Bouzoubaa, "CTNote: CTN Outcomes, Treatments, and Endpoints," 2022, https://CRAN.R-project.org/package=CTNote.

28. National Institute of Minority Health and Health Disparities, "Understanding Health Disparities Series: Diversity & Inclusion in Clinical Trials," 2025, http://web.archive.org/web/20250205203424/, https://www.nimhd.nih.gov/resources/understanding-health-disparities/diversity-and-inclusion-in-clinical-trials.html.

29. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Women in Science, Engineering, and Medicine, Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research, K. Bibbins-Domingo, and A. Helman, "Why Diverse Representation in Clinical Research Matters and the Current State of Representation Within the Clinical Research Ecosystem," in *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups* (National Academies Press (US), 2022).

30. R. E. Johnson, J. H. Jaffe, and P. J. Fudala, "A Controlled Trial of Buprenorphine Treatment for Opioid Dependence," *Journal of the American Medical Association* 267, no. 20 (1992): 2750–2755.

31. W. Ling, C. Charuvastra, J. F. Collins, et al., "Buprenorphine Maintenance Treatment of Opiate Dependence: A Multicenter, Randomized Clinical Trial," *Addiction* 93, no. 4 (1998): 475–486.

32. E. M. Krupitsky, E. E. Zvartau, D. V. Masalov, et al., "Naltrexone for Heroin Dependence Treatment in St. Petersburg, Russia," *Journal of Substance Abuse Treatment* 26, no. 4 (2004): 285–294.

33. T. R. Kosten, R. Schottenfeld, D. Ziedonis, and J. Falcioni, "Buprenorphine Versus Methadone Maintenance for Opioid Dependence," *Journal of Nervous and Mental Disease* 181, no. 6 (1993): 358–364.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** dar70085-sup-0001-Supinfo.docx.