Distributional Diagnosis and Calibration with Negative Controls for Outcome-wide Real-world Evidence

Huiyuan Wang[a,b], Bingyu Zhang[a,c], Yuqing Lei[a,b], Yiwen Lu[a,c], Dazheng Zhang[a,b], Xinyao Jian[a,b], Yuru Zhu[a,b], Wenjie Hu[a,b], Haitao Chu[d], Yong Chen[d], Marc A. Suchard[e,f,g], Patrick B. Ryan[e,h,i], George Hripcsak[e,h], David A. Asch[j,k,l,m,n], Yun Lu[o], Bin Yu[p,q,r], Martijn J. Schuemie[e,f,i], Yumou Qiu*[s], and Yong Chen*[a,b,c,j,t,u]

[a]*The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA*
[b]*Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA*
[c]*The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA*
[d]*Statistical Research and Data Science Center, Pfizer Inc., New York, NY, USA*
[e]*Observational Health Data Sciences and Informatics, New York, NY, USA*
[f]*Department of Biostatistics, University of California, Los Angeles, CA, USA*
[g]*VA Informatics and Computing Infrastructure, US Department of Veterans Affairs, Salt Lake City, UT, USA*
[h]*Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA*
[i]*Global Epidemiology Organization, Johnson & Johnson, Titusville, NJ, USA*
[j]*Leonard Davis Institute of Health Economics, Philadelphia, PA, USA*
[k]*Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[l]*Penn Center for Health Incentives and Behavioral Economics, University of Pennsylvania, Philadelphia, PA, USA*
[m]*Wharton School, University of Pennsylvania, Philadelphia, PA, USA*
[n]*Division of General Internal Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[o]*Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA*
[p]*Department of Statistics, University of California, Berkeley, Berkeley, CA, USA*
[q]*Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA*
[r]*Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA*
[s]*School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China*
[t]*Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA*
[u]*Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA*

Abstract

Glucagon-like peptide-1 receptor agonists (GLP-1RAs) have been linked to heterogeneous, potentially pleiotropic effects across organ systems, motivating outcome-wide comparative risk profiling in real-world data. A central challenge in such analyses is *residual bias* that remains after adjustment for observed confounders, which can distort effect estimates and mis-calibrate uncertainty. We present distributional diagnosis and calibration (DC), which uses panels of negative control outcomes (NCOs) to diagnose residual bias and calibrate uncertainty. DC evaluates null behavior via p-value uniformity and empirical coverage across NCOs, and uses the empirical distribution of NCO effect estimates to calibrate confidence intervals for prespecified primary outcomes. DC is modular: it can wrap around commonly used causal inference methods and operates directly on summary statistics, supporting collaborative research under data-sharing constraints. Using electronic health records from a large U.S. clinical research network (152.7 million patients), we compared GLP-1RAs with sodium–glucose cotransporter 2 inhibitors across 15 prespecified outcomes spanning cardiovascular, mental health, and genitourinary domains using four causal estimators. Across outcomes and methods, DC diagnostics revealed substantial and method-dependent residual systematic error. DC calibration attenuated systematic error signals observed in negative controls and yielded more stable, better-calibrated estimates for clinical outcomes, supporting DC as a practical strategy to strengthen the credibility of outcome-wide real-world CER.

Keywords: Bias calibration, Negative control outcomes, Unmeasured confounding

Disclaimer: The contents are solely the responsibility of the authors and do not necessarily represent the official views of, or an endorsement by, Food and Drug Administration (FDA)/Department of Health and Human Services (HHS) or the U.S. Government.

Introduction

GLP-1 receptor agonists (GLP-1RAs) have reshaped the clinical landscape of type 2 diabetes management. Although initially developed for glycemic control, GLP-1RAs such as semaglutide have demonstrated substantial benefits for weight loss and cardiovascular risk reduction [1]. However, real-world adoption has outpaced the evidence base from randomized controlled trials (RCTs), particularly for secondary outcomes and for patients who differ from trial populations in comorbidity burden, healthcare access, and prescribing context [2, 3].

Observational studies further suggest a pleiotropic profile of GLP-1RAs, with reported associations spanning multiple organ systems, from gastrointestinal adverse events to potential protective signals for substance use disorders and depression [4–6]. This breadth highlights the need to characterize outcome-wide risk–benefit profiles in routine care, where treatment decisions increasingly require balancing effects across cardiovascular, metabolic, neuropsychiatric, and other domains. Large-scale real-world data resources, including electronic health records (EHR) and claims networks, make such systematic evaluation feasible. Yet the same setting poses a distinct inferential challenge: when many outcomes are examined simultaneously, unmeasured confounders can generate correlated spurious signals across the phenome. For example, GLP-1RA users may appear healthier because of differential healthcare engagement, or sicker because of indication-driven prescribing; either pattern can shift multiple endpoints and blur causal effects with systematic bias. This setting motivates a general need for scalable, data-driven tools to diagnose and calibrate residual systematic error in outcome-wide real-world evidence generation.

A fundamental barrier is that counterfactual outcomes are unobserved, so the magnitude and structure of bias cannot be measured directly from the primary endpoint alone. Negative control outcomes (NCOs) — outcomes known not to be causally affected by the treatment — provide a practical lens for detecting hidden bias: any apparent treatment effect on an NCO reflects residual systematic error [7, 8]. Despite their appeal, negative controls are historically used as informal checks [9]. More recent methods that incorporate negative controls into causal estimation often rely on strong structural assumptions or causal restrictions that can be difficult to justify in complex real-world settings [10–12]. Moreover, most of these approaches focus on bias correction but offer limited tools to assess whether meaningful residual bias persists after adjustment, complicating interpretation of real-world evidence. There remains a need for principled, data-driven approaches that leverage *panels* of NCOs to characterize systematic error and support calibrated inference for outcomes whose true effects are unknown.

Here we introduce *distributional diagnosis and calibration* (DC), a unified framework that leverages a panel of NCOs to model and adjust for systematic error in non-interventional studies. Rather than requiring detailed specification of causal structure, DC attributes systematic components of estimation error across outcomes to shared features of the data environment, study design, and analytic pipeline. The key premise is a form of *bias exchangeability*: when the primary outcome and NCOs are analyzed within the same data source using standardized cohort construction, covariate definitions, and a common estimation procedure, residual confounding and model misspecification tend to induce systematic error components with similar behavior across outcomes. By learning the empirical distribution of estimated NCO effects, DC provides an internal diagnostic of residual bias and yields calibrated uncertainty for the primary endpoint.

We demonstrate DC using data from the TriNetX Research Network, a global EHR platform aggregating de-identified patient data from hundreds of healthcare organizations. To characterize residual bias in this setting, we leverage a prespecified panel of 95 NCOs expected to have null treatment effects based on clinical knowledge. We perform leave-one-out analyses, treating each NCO in turn as a pseudo-primary endpoint, to map patterns of spurious association and quantify the magnitude and dispersion of systematic error under known null conditions. Guided by these empirically observed bias patterns, we estimate and calibrate comparative treatment effects of GLP-1RAs versus sodium–glucose cotransporter 2 inhibitors (SGLT2is) across the 15 prespecified outcomes. Together, these results illustrate how DC provides an internal,

data-driven assessment of residual systematic error and yields calibrated uncertainty for outcome-wide real-world comparative effectiveness analyses.

Results

Overview of DC

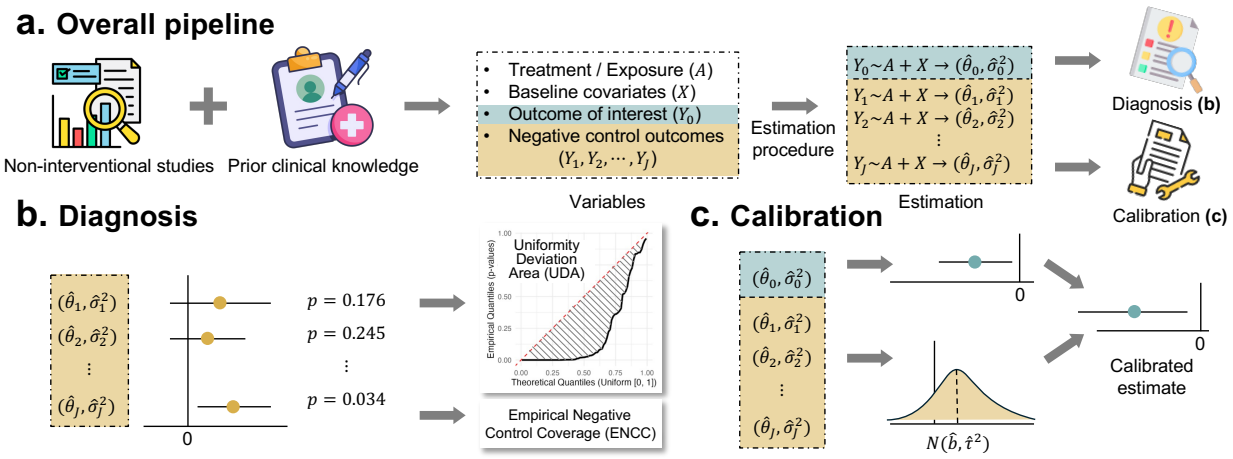

Figure 1: **Overview of distributional diagnosis and calibration using negative control outcomes. a.** Study setup: integrate real-world data and prior clinical knowledge to define treatment (A), covariates (X), primary outcome (Y_0), and a prespecified panel of negative control outcomes ($Y_1, \ldots, Y_J$), followed by model-based estimation. **b.** Diagnosis: use the negative-control panel to characterize residual systematic error (e.g., QQ plots and empirical null coverage). **c.** Calibration: use the empirical distribution of negative-control estimates to calibrate the primary estimate and its uncertainty.

Figure 1 summarizes distributional diagnosis and calibration (DC), a post-estimation module for diagnosing and calibrating residual systematic error in observational causal analyses. DC takes as input the point estimates and standard errors produced by a prespecified estimation pipeline and returns (i) diagnostics that characterize whether negative controls behave as expected under the null and (ii) calibrated inference for the prespecified clinical outcomes. Conceptually, DC consists of a setup step, a diagnostic layer based on a negative-control panel, and a calibration step that propagates empirically observed systematic error into the primary inference.

Diagnostic layer (b). DC uses the negative-control panel as an exploratory lens to map residual bias patterns induced by the shared data source and analytic pipeline. First, we examine quantile–quantile (QQ) plots of negative-control p-values. If residual systematic error is negligible and uncertainty is well-calibrated among the negative control outcomes whose true treatment effects are expected to be null, the p-values should be approximately uniform and the QQ curve should track the 45-degree reference line; systematic departures suggest residual confounding and/or mis-calibrated uncertainty. We summarize the deviation using the Uniformity Deviation Area (UDA), defined as the area between the empirical QQ curve and the 45-degree line, with larger values indicating stronger evidence of systematic error. Second, Empirical Negative Control Coverage (ENCC) quantifies how often nominal confidence intervals for negative controls include zero. Coverage below the nominal level indicates remaining bias and/or underestimated uncertainty.

Calibration layer (c). DC calibrates inference using the empirical distribution of negative-control estimates. In brief, DC recenters the primary estimate by the average estimated negative-control effect and inflates uncertainty to reflect dispersion across negative controls, thereby accounting for both systematic shift and over/under-dispersion. This calibration is adaptive: when negative controls show little evidence of systematic error, DC makes minimal changes; when negative controls indicate substantial residual bias, DC applies a stronger correction and widens uncertainty accordingly. Under mild regularity conditions, the resulting confidence intervals achieve valid coverage.

Application to outcome-wide comparative effectiveness in TriNetX. We applied DC to the TriNetX Research Network, a global EHR platform aggregating de-identified patient data from hundreds of healthcare organizations. The clinical objective was to compare GLP-1RAs versus SGLT2is across 15 prespecified outcomes spanning cardiovascular, mental health, and genitourinary domains. Each outcome was analyzed as an incident event using an outcome-specific washout procedure, yielding outcome-specific analytic cohorts. We implemented DC with four causal estimators and applied a common prespecified negative-control panel to characterize residual systematic error in this data environment.

Across the 15 outcomes, DC diagnostics provided an internal assessment of residual systematic error under each estimation pipeline, and DC calibration yielded calibrated uncertainty for comparative effect estimates. In the main text, we report the calibrated outcome-wide results and the corresponding negative-control diagnostics; additional results under alternative estimators and extended analyses are provided in the *Supplementary Information*.

DC diagnostics reveal residual systematic error across estimation methods

We assessed residual systematic error across combinations of estimation method and covariate adjustment. In each setting, we applied DC diagnostics to 95 prespecified NCOs, which are expected to have null treatment effects. This design provides an internal check of whether the analysis pipeline yields well-calibrated null behavior in the same data environment as the clinical outcomes.

QQ plots and UDA for uncalibrated estimates. Figure 2 shows substantial departures from Uniform$(0, 1)$ behavior for uncalibrated NCO p-values (gray curves) across estimation methods and adjustment levels, indicating residual systematic error and/or mis-calibrated uncertainty that persists after standard confounding adjustment. Increasing covariate adjustment generally moved the p-value distributions closer to the 45-degree line, but noticeable deviations remained even under extensive adjustment (up to $d = 178$ covariates), suggesting that richer adjustment alone did not fully restore well-calibrated null behavior.

To summarize these departures, we computed the Uniformity Deviation Area (UDA). Table 1 reports UDA values together with empirical null quantiles. For each estimator, UDA decreases systematically with more extensive covariate adjustment, indicating improved (but still imperfect) alignment with the expected null behavior. The Q95 and Q99 thresholds are obtained from a parametric bootstrap null that preserves the observed covariate structure by re-simulating treatment and outcomes from fitted models (excluding unmeasured confounding under the null); implementation details are provided in *Methods*.

Empirical coverage for uncalibrated estimates. We further assessed uncertainty calibration by examining the empirical coverage of nominal 95% confidence intervals across the 95 NCOs. Specifically, we drew 200 random subsamples from the analytic cohort; within each subsample, we applied the prespecified causal estimator to each of the 95 NCOs to obtain a nominal 95% confidence interval for its estimated treatment effect, and then computed the fraction of these intervals that contained zero. This yields one empirical

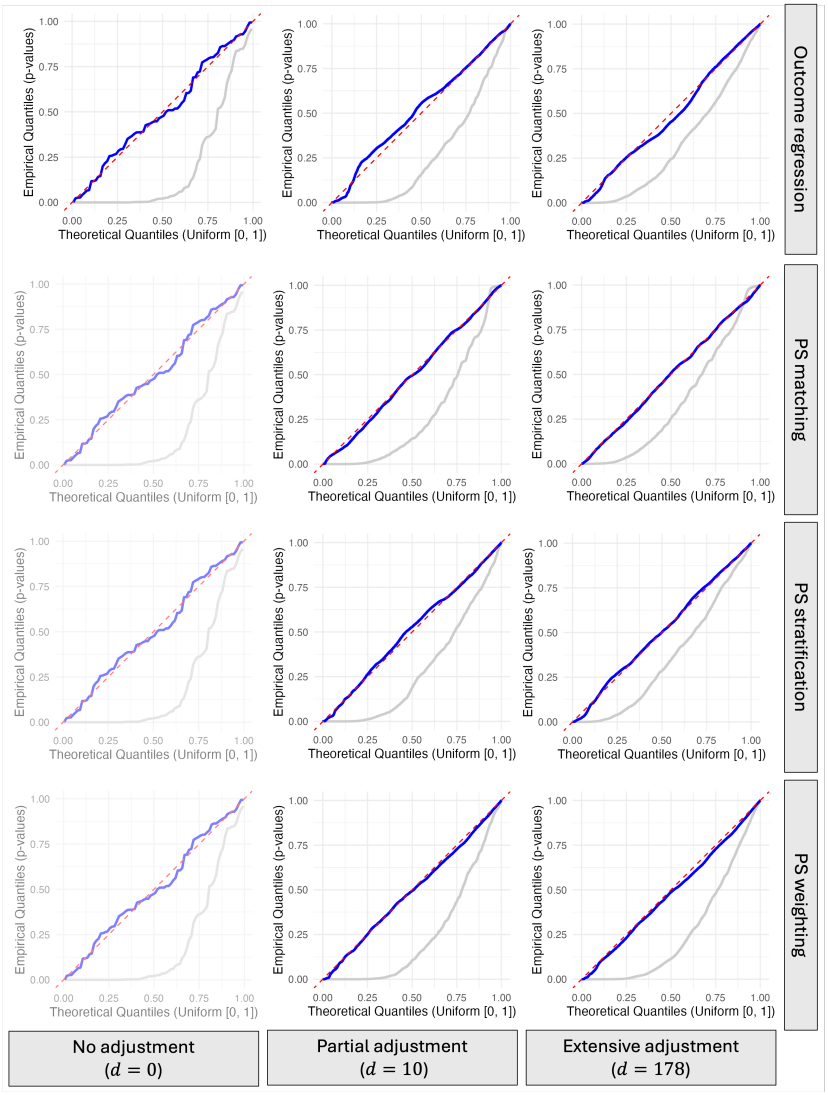

Figure 2: **DC diagnostics before and after calibration.** Each panel shows a QQ plot comparing empirical p-values from 95 NCOs to the Uniform$(0, 1)$ reference across covariate adjustment levels (columns) and estimation methods (rows). Gray curves show uncalibrated p-values; blue curves show calibrated p-values obtained by leave-one-out DC using the NCO panel. The number of adjusted covariates is denoted by d (the no-adjustment column is identical across methods). The red dashed line indicates ideal uniformity.

coverage value per subsample. Figure 3 summarizes the resulting distribution of coverage values across subsamples.

Across estimation methods and covariate adjustment levels, uncalibrated procedures (faint gray boxplots) frequently fell below the nominal 95% level, indicating mis-calibrated uncertainty and/or residual systematic error. Although more extensive covariate adjustment tended to improve coverage on average, coverage remained variable even under extensive adjustment, reflecting sensitivity to sampling variation and persistent systematic error in the observational pipeline.

Together with the QQ-based diagnostics, these results show that standard covariate adjustment alone often does not restore well-calibrated null behavior, motivating post-estimation calibration.

Table 1: **Quantitative diagnostics for p-value uniformity.** We report UDA across estimation methods and covariate adjustment levels, before and after DC calibration. For each setting, Q95 and Q99 denote the 95th and 99th percentiles of the UDA obtained by a parametric bootstrap procedure; UDA values exceeding these thresholds indicate departures from Uniform$(0, 1)$ behavior. Across settings, uncalibrated UDAs substantially exceeded the empirical null thresholds, whereas DC calibration consistently reduced UDA below the corresponding Q95 benchmark, indicating markedly improved alignment with the expected null behavior.

Adjustment Level	Method	Calibration	Area	Q95	Q99
No adjustment ($d = 0$)	None (Unadjusted)	Uncalibrated	0.303	0.062	0.073
		Calibrated	0.023	0.062	0.073
Partial adjustment ($d = 10$)	PS matching	Uncalibrated	0.225	0.075	0.092
		Calibrated	0.028	0.075	0.092
	PS stratification	Uncalibrated	0.226	0.083	0.109
		Calibrated	0.020	0.083	0.109
	PS weighting	Uncalibrated	0.251	0.080	0.104
		Calibrated	0.016	0.080	0.104
	Outcome regression	Uncalibrated	0.245	0.137	0.154
		Calibrated	0.027	0.137	0.154
Extensive adjustment ($d = 178$)	PS matching	Uncalibrated	0.181	0.075	0.095
		Calibrated	0.022	0.075	0.095
	PS stratification	Uncalibrated	0.161	0.072	0.091
		Calibrated	0.014	0.072	0.091
	PS weighting	Uncalibrated	0.232	0.082	0.104
		Calibrated	0.021	0.082	0.104
	Outcome regression	Uncalibrated	0.180	0.090	0.110
		Calibrated	0.027	0.090	0.110

DC calibration restores well-calibrated null behavior across estimation methods

We next evaluated whether DC yields well-calibrated confidence intervals and hypothesis tests in the presence of residual systematic error. Because the same negative-control panel informs both diagnosis and calibration, we used a leave-one-out (LOO) strategy to avoid reusing the same NCO for fitting and evaluation. In each iteration, one NCO was held out, the remaining NCOs were used to estimate the empirical bias distribution, and the held-out NCO was then calibrated using this distribution (see *Methods*).

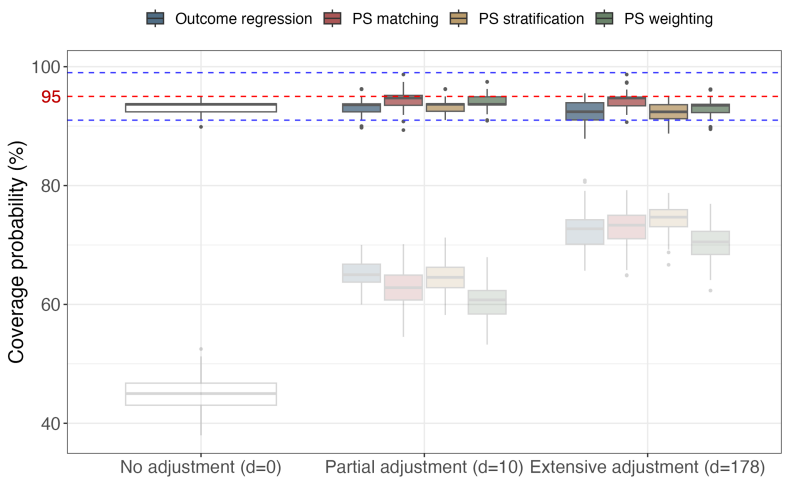

Figure 3: **Empirical negative-control coverage before and after calibration.** For each subsample and analysis setting, we applied LOO-based DC calibration to obtain calibrated 95% confidence intervals for the treatment effects of all NCOs, and computed the fraction of NCO intervals that contained zero (one coverage value per subsample). Boxplots summarize coverage across 200 random subsamples, stratified by covariate adjustment level (x-axis) and estimation method (color). The red dashed line marks nominal 95% coverage; blue dashed lines indicate 91% and 99% reference bounds. Faint boxplots show uncalibrated coverage for comparison (reproduced from Fig. S.3 in the Supplementary Material).

Empirical coverage after calibration. To assess stability across sampling variation, we repeated the LOO procedure across 200 random subsamples and computed an empirical null-coverage value in each subsample. Figure 3 shows that DC calibration brings empirical coverage closer to the nominal 95% level across estimation methods and adjustment levels, while substantially reducing between-subsample variability relative to uncalibrated analyses. The largest improvements occur in under-adjusted settings, consistent with DC correcting both systematic shift and underestimation of uncertainty under residual bias.

Consistency with QQ-based diagnostics. The coverage improvements are consistent with the QQ-based diagnostics: in Fig. 2, LOO-based calibration moves negative-control p-value distributions closer to Uniform$(0, 1)$ across methods and adjustment levels. Likewise, Table 1 shows that calibrated UDA values fall below the corresponding empirical null Q95 thresholds across settings, indicating markedly improved null behavior after calibration.

DC calibration gains efficiency with stronger covariate adjustment

Having established that DC improves uncertainty calibration under residual systematic error, we next examined how covariate adjustment affects the precision of calibrated inference. DC widens confidence intervals to reflect uncertainty induced by residual systematic error estimated from the negative-control panel, which raises a practical question: can stronger covariate adjustment reduce the amount of inflation required for calibration?

To address this question, we applied LOO-based DC to the 95 NCOs under increasing levels of covariate adjustment while holding the estimator fixed. At each adjustment level, we computed the mean length of calibrated and uncalibrated 95% confidence intervals across NCOs and summarized results over 200 random

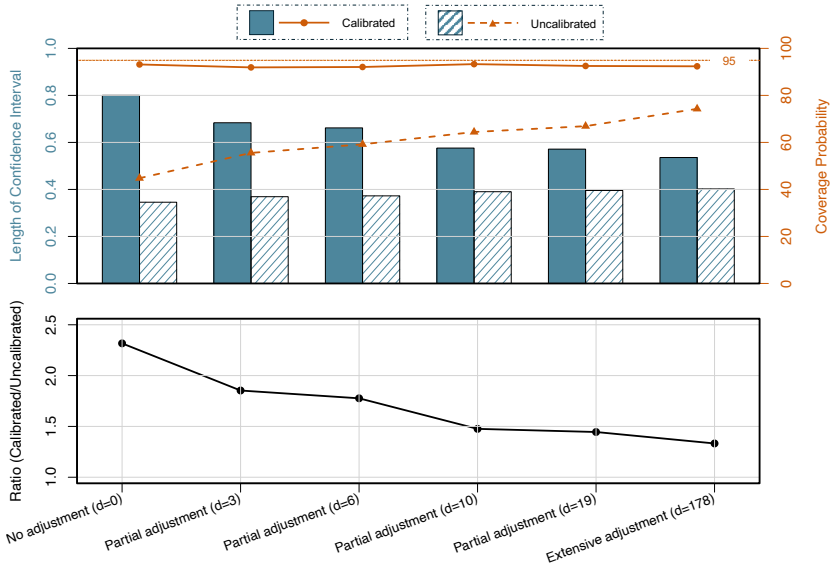

Figure 4: **Precision and coverage under DC calibration across covariate adjustment levels.** Across covariate adjustment levels, we summarize (a) the mean length of nominal 95% confidence intervals (CIs) before and after LOO-based DC calibration, (b) the corresponding empirical coverage for negative controls, and (c) the ratio of calibrated to uncalibrated CI lengths. Curves and boxplots summarize results across 200 random subsamples.

subsamples. We report results for propensity score stratification in the main text; results for propensity score weighting, propensity score matching, and outcome regression are provided in the *Supplementary Materials*.

Figure 4 shows a clear efficiency gain with stronger adjustment. As more covariates are included, calibrated intervals become shorter on average, and the ratio of calibrated to uncalibrated interval lengths decreases, indicating that the additional uncertainty propagated by DC is substantially reduced under richer adjustment. At the same time, empirical null coverage remains close to the nominal 95% level after calibration.

These results highlight a complementary relationship between adjustment and calibration. Covariate adjustment reduces residual systematic error and shrinks the dispersion of negative-control estimates, which in turn lowers the uncertainty that DC must propagate during calibration. DC then accounts for the remaining systematic error and yields calibrated inference, with the efficiency of the final estimates improving as pre-calibration adjustment becomes stronger.

Comparative effectiveness across cardiovascular, mental health, and genitourinary outcomes

GLP-1RAs and SGLT2is are widely used glucose-lowering therapies for individuals with type 2 diabetes, and both have established cardiometabolic benefits in randomized trials. To characterize their comparative outcome profiles across multiple clinically relevant domains, we applied DC calibration to 15 incident outcomes spanning cardiovascular (acute hemorrhagic cerebrovascular disease, acute phlebitis/thrombophlebitis and thromboembolism, acute pulmonary embolism, cerebral infarction, hypotension, nonspecific chest pain), mental health (alcohol-related disorders, anxiety and fear-related disorders, depressive disorders, neurodevelopmental disorders, opioid-related disorders), and genitourinary conditions (acute and unspecified renal failure, chronic kidney disease, erectile dysfunction, menstrual disorders). Each outcome was defined using a 12-month washout period to capture new-onset events.

For effect estimation, we used propensity score stratification with the full set of prespecified covariates (extensive covariate adjustment), followed by DC calibration using the prespecified negative-control panel.

Figure 5 summarizes the calibrated log risk ratios (logRRs) and 95% confidence intervals for all 15 outcomes. Across cardiovascular endpoints, calibrated estimates were generally close to the null and confidence intervals included no clear differences between GLP-1RAs and SGLT2is. For example, the calibrated logRR for acute pulmonary embolism was -0.21 (95% CI, -0.86 to 0.44), consistent with similar risk between therapies in this setting.

Mental health outcomes showed greater heterogeneity and wider uncertainty. Most calibrated estimates remained imprecise and crossed the null, indicating limited comparative evidence across these domains in the available data. Neurodevelopmental disorders showed an elevated calibrated estimate for GLP-1RA initiation (logRR 0.77; 95% CI, 0.11 to 1.44); this finding should be interpreted cautiously given the low event rate and the possibility of residual systematic error that may differ across outcome definitions and subpopulations.

For genitourinary outcomes, menstrual disorders showed a positive calibrated association (logRR 0.91; 95% CI, 0.25 to 1.58), whereas erectile dysfunction had a positive point estimate with substantial uncertainty (logRR 0.22; 95% CI, -0.39 to 0.83).

Overall, these DC-calibrated results provide a bias-aware summary of comparative risks across the prespecified outcome set by propagating empirically observed systematic error from negative controls into uncertainty for the clinical outcomes. This calibration supports more transparent interpretation of outcome-wide real-world comparative effectiveness analyses.

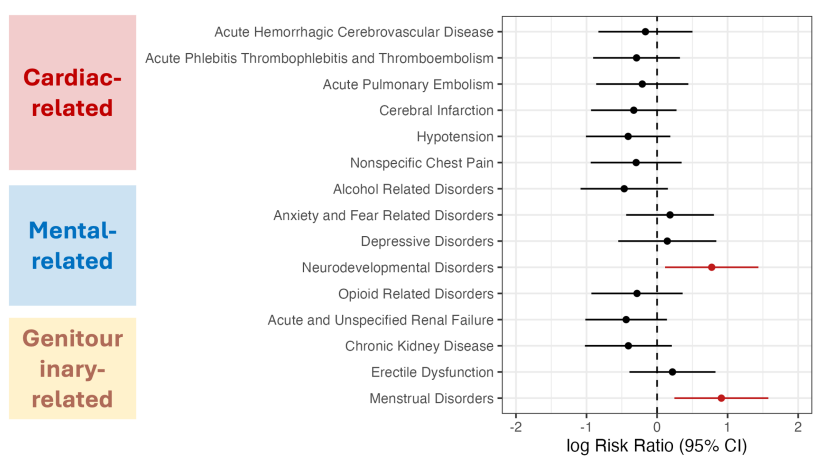

Figure 5: **DC-calibrated comparative estimates across 15 prespecified incident outcomes.** Shown are calibrated log risk ratios (logRRs) and 95% confidence intervals comparing GLP-1RAs with SGLT2is across cardiovascular, mental health, and genitourinary domains. Red points indicate outcomes for which the calibrated 95% confidence interval excludes the null (logRR $= 0$).

Discussion

We developed distributional diagnosis and calibration, a framework that uses panels of negative control outcomes to diagnose and calibrate residual systematic error in real-world comparative effectiveness analyses. In the empirical evaluation on a large-scale EHR dataset, DC diagnostics indicated that commonly used adjustment strategies can leave substantial residual systematic error, as reflected by departures from expected null behavior and under-coverage of nominal confidence intervals, particularly under limited covariate adjustment. Applying DC calibration improved alignment with null behavior (e.g., p-value uniformity and empirical coverage) across estimation methods. We also found that richer covariate adjustment improved the efficiency of calibrated inference by shrinking the dispersion of NCO effects, thereby reducing the amount of

uncertainty that DC must propagate. Together, these results suggest that covariate adjustment and calibration play complementary roles: adjustment reduces systematic error at its source, and DC provides an internal, data-driven safeguard when residual error persists.

A key feature of DC is its *flexibility and modularity*. Because it operates on point estimates and standard errors, DC can be wrapped around a wide range of causal inference pipelines without altering the underlying estimation procedure. This design supports interpretability and reproducibility and makes DC broadly applicable to non-interventional study designs, including observational cohort and case-control studies, self-controlled case series (SCCS), self-controlled risk interval (SCRI) designs, and case-crossover analyses.

DC is particularly relevant for high-stakes settings in regulatory science, pharmacoepidemiology, and public health surveillance, where decisions rely on credible uncertainty quantification and where spurious signals can have substantial downstream consequences [13, 14]. By calibrating inference using an internal empirical null constructed from negative controls, DC emphasizes reliable uncertainty assessment. In practice, associations that remain stable after calibration may be viewed as more robust, whereas estimates that shift toward the null or acquire substantially wider uncertainty warrant more cautious interpretation. More broadly, DC provides a transparent and scalable approach to strengthen real-world evidence pipelines when residual systematic error is difficult to rule out.

Several limitations warrant further study. First, our current calibration models the distribution of NCO-based bias estimates using a parametric form. While this approximation is supported empirically in our analyses and can be justified under mild conditions (Supplementary Information), more flexible or assumption-lean calibration strategies may further improve robustness. Second, DC depends on the validity and representativeness of the NCO panel: violations such as latent causal pathways or outcome misclassification could degrade performance, motivating robust procedures that detect and downweight potentially invalid controls. Third, constructing high-quality NCO panels currently relies on domain expertise; future work could integrate semi-automated candidate generation using biomedical knowledge resources (e.g., SemMedDB [15].) with clinical review to improve scalability in high-throughput applications.

Overall, DC operationalizes a general principle: leverage prior causal knowledge encoded by negative controls to learn about systematic error affecting outcomes of interest. This complements existing causal pipelines by enabling data-driven diagnosis and calibrated uncertainty without requiring explicit modeling of the full data-generating process. We view DC as a practical step toward improving the reliability and interpretability of outcome-wide real-world evidence, particularly in settings where rigorous error control is essential.

Methods

Data source and cohort construction

We used data from the TriNetX Research Collaborative Network. This network provides access to de-identified, longitudinal electronic medical records from approximately 152.7 million patients across 106 healthcare organizations, including detailed information on diagnoses, procedures, medications, laboratory values, and genomic data. The data are accessible via the TriNetX Analytics Network (https://trinetx.com). The use of TriNetX for this study was reviewed and approved by the *Institutional Review Board* of the University of Pennsylvania #858496.

Target clinical question and outcome. Our clinical objective was to evaluate the comparative effectiveness of GLP-1RAs versus SGLT2is of cardiovascular, mental health, and genitourinary outcomes. Detailed eligibility criteria are provided in *Methods*, and the specific medications included in each treatment group are listed in the *Supplementary Materials*.

Eligibility criteria. The study period spanned from January 1, 2019 to September 30, 2024. We included individuals aged 18 years or older who initiated a GLP-1RA or SGLT2i during this period, with no prescriptions for either medication in the previous 12 months. The index date was defined as the date of first prescription for either drug. Patients were required to have at least one healthcare encounter (outpatient, inpatient, or emergency department) in the 12 months prior to the index date. Those who initiated both treatments on the same day were excluded. Participants with specific outcomes during the baseline period (12 month prior to index date) were excluded.

NCO selection. A pre-specified list of 95 NCOs, adapted from existing literature [16], was used to evaluate and correct for residual bias and other sources of systematic error. Details can be found in *Supplementary Materials*.

Covariate selection. Patient covariates were selected to account for potential confounding which included demographic variables (age, sex, race/ethnicity), clinical characteristics (e.g., baseline comorbidities such as hypertension, obesity, and type 2 diabetes mellitus (T2DM)), and year of cohort entry (ranging from January 2019 to September 2024). These variables were treated as confounders and selected based on clinical relevance and prior literature. However, it should be noted that, as a key challenge in EHR-based studies, not all relevant confounders are routinely or reliably captured, making the presence of unmeasured confounding a persistent concern.

Causal estimation methods deployment

Estimation methods. To adjust for measured confounding, we employed four commonly used estimation methods: propensity score (PS) matching, PS stratification, inverse probability of treatment weighting (IPTW), and multivariable regression with direct covariate adjustment. These approaches enable a comparison of calibration performance across different adjustment paradigms.

Adjustment levels. To assess how covariate inclusion affects bias calibration, we defined three levels of adjustment, where d is the number of covariates included: (1) no adjustment, with no covariates included ($d = 0$); (2) partial adjustment, using nested subsets of clinically relevant covariates ($d = 3, 6, 10, 19$); and (3) extensive adjustment, incorporating all available demographic and clinical covariates ($d = 178$). Further details are provided in *Methods*.

Evaluation dimensions. We evaluated the impact of different covariate adjustment strategies and the proposed distributional calibration procedure across three dimensions: diagnostic performance before and after calibration, and estimation efficiency.

Causal estimation methods. To mitigate the effects of observed confounders, we applied four commonly used methods: propensity score (PS) matching, PS stratification, PS inverse probability of treatment weighting (IPTW), and direct covariate adjustment via regression.

The first three methods rely on a two-step procedure. In the first step, we estimated the propensity score using logistic linear regression, modeling the probability of receiving treatment A as a function of the observed covariates X. In the second step, we estimated treatment effects using outcome regression ($Y \sim A$) after adjusting the observed covariates by each method using the estimated propensity scores. A modified Poisson linear regression model [17] was used due to its favorable properties for estimating risk ratios [18].

For PS matching, individuals in the treatment and control groups were matched in a 1:1 ratio, and the outcome model was applied to the matched cohort. For PS stratification, individuals were grouped into

quintiles of the estimated PS, and a conditional Poisson regression adjusting for the PS strata was used. For PS weighting, we used stabilized weights based on the inverse of the estimated PS (for treated) or one minus the PS (for controls), and we estimated treatment effects via a weighted Poisson regression. Lastly, for outcome regression adjustment, we directly fit a multivariate modified Poisson regression model including all observed covariates.

Selection of partially adjusted covariates. To evaluate the robustness of distributional calibration under varying levels of confounding adjustment, we predefined four nested tiers of partial covariate adjustment. Each tier incrementally incorporated clinically meaningful variables selected based on domain expertise and prior literature, reflecting potential confounders of the treatment-outcome relationship.

- Demographics only (3 covariates): Basic demographic covariates, including age, sex, and race/ethnicity.
- Demographics and Core Metabolic Comorbidities (6 covariates): Demographics plus key metabolic comorbidities, including age, sex, and race/ethnicity, type 2 diabetes mellitus, obesity, and hypertension.
- Core and Cardiovascular Conditions (10 covariates): All covariates above, plus additional cardiometabolic comorbidities, including disorders of lipid metabolism, heart failure, cardiac arrhythmias, cardiorespiratory signs and symptoms
- More Clinical Profile (19 covariates): The most comprehensive partial set, further including sleep-wake disorders, anxiety disorder, general signs and symptoms, chest pain, chronic kidney disease, abdominal pain, low back pain, headache, and asthma.

The extensive adjustment list included 178 covariates, encompassing demographics (age, sex, race/ethnicity), detailed clinical characteristics, and the calendar year of cohort entry (ranging from January 2019 to September 2024). The codeset for clinical conditions is available at `https://github.com/Penncil/Distributional-diagnosis-and-calibration/blob/master/cluster_master.csv`, and conditions with a prevalence below 0.5% were excluded to enhance computational efficiency.

Statistical notation

We first introduce necessary notation. We adopt the potential outcomes framework [19], where $A \in \{0, 1\}$ denotes a binary treatment, X the observed covariates, and U unmeasured confounders. For each outcome Y_j, let $Y_j(1)$ and $Y_j(0)$ denote the potential outcomes under treatment and control, respectively, with $Y_j = AY_j(1) + (1 - A)Y_j(0)$. We index the primary outcome as Y_0, and define Y_j with $j \geq 1$ as an NCO if $\mathbb{E}\{Y_j(1)\} = \mathbb{E}\{Y_j(0)\}$. We observe n independently and identically distributed (i.i.d.) samples of $(Y_{i,0}, A_i, X_i, Y_{i,1}, \ldots, Y_{i,J}), i = 1, \ldots, n$. See Figure 1 for illustration.

Let β_A denote a generic notation of the target causal estimand for the primary outcome Y. The causal parameter β_A may vary depending on context and application. Common choices include:

- **The average treatment effect**: $\beta_{\text{ATE}} := \mathbb{E}\{Y(1) - Y(0)\}$;
- **The log risk ratio**: $\beta_{\text{logRR}} := \log\left\{\mathbb{E}\{Y(1)\}/\mathbb{E}\{Y(0)\}\right\}$.

For a given estimation method of β_A, let $\hat{\theta}_j$ denote the estimated treatment effect for the jth outcome/NCO, where $j = 0$ corresponds to the outcome of interest and $j = 1, \ldots, J$ index a set of NCOs. Let $\theta_j^\star$ denote the asymptotic limit of $\hat{\theta}_j$, which represents the value of the estimate by the estimation procedure as if we had infinitely many samples. Note that $\theta_0^\star$ may not equal β_A and $\theta_j^\star$ may not equal 0 for NCOs due to systematic error. In the main text, we focus on the causal parameter ATE, namely $\beta_A = \beta_{\text{ATE}}$, though the methodology generalizes to other estimands.

Framework overview: distributional diagnosis and calibration with NCOs

Importantly, the framework does not require specification of a data-generating model or causal directed acyclic graphs, making it broadly applicable across settings. See *Supplementary Materials* (Examples S.1–S.3) for detailed discussions of its use under different structural assumptions.

DC framework layer (a) — setup stage: integrating NCOs into the causal estimation pipeline

In comparative effectiveness research, it is a standard practice to follow a structured protocol: define the scientific question, select the eligible cohort, identify relevant covariates, and estimate treatment effects using adjustment strategies such as propensity score matching. Our framework extends this protocol by incorporating NCOs for bias diagnosis and correction. The setup proceeds as follows:

1. *Define the causal question and eligible population.* Specify the treatment variable A and primary outcome Y_0 based on the causal estimand of interest. Define the eligible population from the target real-world database using prespecified inclusion and exclusion criteria.

2. *Select key covariates and identify NCOs.* Use prior scientific and clinical knowledge to identify potential confounders and NCOs. Specifically, NCOs are outcomes $Y_1, \ldots, Y_J$ that are known to be causally unaffected by A but share a similar confounding structure with the primary outcome.

3. *Estimate treatment effects.* Apply the chosen estimation method to each outcome to obtain a collection of point estimates and associated variance estimates: $\{(\hat{\theta}_j, \hat{\sigma}_j^2)\}_{j=0}^{J}$, where $j = 0$ corresponds to the primary outcome and $j = 1, \ldots, J$ index the NCOs.

This setup stage establishes the foundation for the subsequent diagnostic and calibration procedures. A schematic illustration of this setup is provided in panel (a) of Figure 1.

DC framework layer (b) — diagnostic layer: detecting residual bias

We illustrate the proposed diagnosis method in panel (b) of Figure 1. Given an estimation procedure, we assess whether it sufficiently removes the residual confounding bias using the estimates $\{\hat{\theta}_j\}$ and standard errors $\{\hat{\sigma}_j\}$ from the NCOs. We propose two complementary diagnostic strategies: (i) the uniformity deviation area test, and (ii) empirical negative control coverage evaluation.

Quantile-quantile plot and uniformity deviation area. Under the null hypothesis that the estimator is unbiased, treatment effects on NCOs should be centered at zero, and the corresponding p-values should follow a Uniform$(0, 1)$ distribution. We visualize this behavior using Quantile-Quantile (QQ) plots and summarize deviations from the uniformity line (the 45-degree diagonal) by the total area between the QQ curve and the diagonal. We term this quantity the *Uniformity Deviation Area* (UDA), which serves as a scalar diagnostic for systematic bias. Formally, UDA can be used to test the hypothesis:

$$H_0 : \text{the causal estimator is unbiased across all NCOs.}$$

Using the parametric bootstrap, we derive the empirical distribution of UDA under H_0 in *Methods*, enabling principled statistical inference. Since the p-values for NCOs are computed on the same dataset, they are generally correlated, violating the independence assumptions required by classical asymptotic theory. This dependence structure motivates the use of the parametric bootstrap, which accounts for such correlations and yields more accurate empirical quantiles. In practice, we estimate the 95th and 99th percentiles of this null distribution using $B = 1{,}000$ bootstrap replicates to ensure stable quantile estimation. If the observed

UDA falls below these critical values, we retain H_0 and interpret the result as consistent with unbiased estimation. Conversely, if the observed UDA exceeds these thresholds, we have statistical evidence to reject H_0, suggesting the presence of systematic bias in the estimation procedure.

Empirical coverage probability. As a complementary diagnostic, we evaluate the empirical coverage of the confidence intervals constructed for NCOs. Since the true treatment effect on each NCO is known to be zero based on prior knowledge, correct coverage requires that zero falls within the nominal $(1-\alpha)$ confidence interval at the expected frequency. For each NCO, we record whether its confidence interval covers zero, and average across J outcomes to estimate empirical coverage. We refer to this procedure as the *Empirical negative control coverage* (ENCC).

Since the empirical coverage is itself a random quantity based on a finite number of NCOs, it naturally fluctuates around the nominal level. For instance, with J NCOs and a nominal level of $1-\alpha = 0.95$, the empirical coverage under the null follows a Binomial distribution $\text{Bin}(J, 0.95)$. Under the central limit theorem, an approximate 95% confidence interval for the empirical coverage can be computed as $[0.95 \pm 1.96\{0.95 \times 0.05 J^{-1}\}^{1/2}]$, which evaluates to approximately $[0.91,\ 0.99]$ when $J = 95$. A significant drop below this range suggests residual bias or misestimated variance.

DC framework layer (c)—calibration layer: correcting for residual bias

Panel (c) of Figure 1 is a schematic illustration of the distributional calibration. The diagnostic and calibration layers are implemented as a unified two-stage procedure. In particular, when the diagnostics suggest substantial bias, the calibration has a greater effect, whereas when residual bias is negligible, the calibration leaves the primary estimate largely unchanged. This is done by modeling the distribution of biases implied by the NCO estimates. Specifically, we compute $(\hat{b}, \hat{\tau}^2)$, where $\hat{b}$ denotes the average estimate of NCO causal effects, and $\hat{\tau}^2$ denotes the variance of estimated NCO causal effects. These quantities are then used to construct a bias-corrected estimator and adjusted variance for the primary outcome:

$$\hat{\theta}_0^{\text{cal}} = \hat{\theta}_0 - \hat{b}, \qquad \widehat{\text{Var}}(\hat{\theta}_0^{\text{cal}}) = \hat{\sigma}_0^2 + \hat{\tau}^2.$$

The resulting calibrated confidence interval is given by:

$$\hat{\theta}_0^{\text{cal}} \pm z_{1-\alpha/2} \left\{ \widehat{\text{Var}}(\hat{\theta}_0^{\text{cal}}) \right\}^{1/2},$$

which is provably valid under mild regularity conditions. Here, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ standard normal quantile. Further technical details are provided in *Methods*.

LOO-based DC: evaluating DC through the leave-one-out strategy. In the real-world example, the same set of NCOs is used for both diagnosis and calibration. We implement a leave-one-out (LOO) strategy to prevent circular inference. Specifically, in each iteration, one NCO is held out, and the remaining $J-1$ NCOs are used to estimate the residual bias distribution. Let $(\hat{b}_{-j}, \hat{\tau}^2_{-j})$ denote the estimated mean and variance of residual bias computed from the $J-1$ NCOs excluding the jth one. Here, $\hat{b}_{-j}$ captures the average residual bias and $\hat{\tau}^2_{-j}$ reflects its variability. These quantities are then used to calibrate the treatment effect estimate and variance for the held-out NCO:

$$\hat{\theta}_j^{\text{cal}} = \hat{\theta}_j - \hat{b}_{-j}, \qquad \widehat{\text{Var}}(\hat{\theta}_j^{\text{cal}}) = \hat{\sigma}_j^2 + \hat{\tau}^2_{-j},$$

where $\hat{\theta}_j$ denotes the point estimate for the jth NCO produced by the chosen estimation method, and $\hat{\sigma}_j^2$ is its corresponding model-based variance estimate. We refer to this procedure as the LOO-powered distributional calibration.

We then apply the UDA and ENCC diagnostics to the cross-validated collection $\{(\hat{\theta}_j^{\text{cal}}, \widehat{\text{Var}}(\hat{\theta}_j^{\text{cal}}))\}_{j=1}^{J}$. Specifically, these calibrated estimates are used to compute p-values for uniformity assessment and to evaluate empirical coverage after calibration. This design ensures that diagnostic evaluations reflect the performance of the bias-corrected procedure itself, free from circularity or self-referential inflation. We refer to this diagnostic procedure as the LOO-based DC.

Framework for DC

Due to unmeasured confounding, the asymptotic estimand $\theta_0^\star$ by the chosen estimation method for the outcome of interest may differ from β_A, and we write

$$\theta_0^\star = \beta_A + \text{bias}. \tag{1}$$

Since β_A is not directly identifiable from observed data, we leverage NCOs to infer the residual bias. The key idea is that the true causal effect for each NCO is zero; hence, their asymptotic estimates $\theta_1^\star, \ldots, \theta_J^\star$ reflect residual biases due to unmeasured confounders rather than causal relationships. Formally, we make the following assumption for identification of β_A.

Assumption 1. *The asymptotic biases for NCOs follow a common distribution independently,*

$$\theta_1^\star, \ldots, \theta_J^\star \sim Q,$$

and that the residual bias in the primary outcome is exchangeable with those of the NCOs: $\theta_0^\star - \beta_A \sim Q$. Here, Q represents a latent distribution that governs the bias structure across outcomes.

Assumption 1 formalizes the idea of distributional exchangeability, stating that the residual bias in the target outcome is drawn from the same distribution as that of the NCOs. While this assumption is not empirically verifiable, they can be justified by design: for instance, when NCOs are selected to mirror the confounding structure of the outcome of interest, their bias distributions are expected to follow a common generative process. In *Supplementary Materials*, we give interpretations for Q under linear models. Moreover, the assumption of distributional exchangeability is notably flexible. It does not require the underlying confounding mechanisms to be identical across outcomes, only that the induced biases are governed by a shared probabilistic model. This mild assumption enables the method to accommodate diverse forms of causal directed acyclic graphs (DAG) commonly in real-world settings. See Examples S.1–S.3 for more details.

Diagnosis for residual bias: uniformity area deviation test

To formally assess whether a candidate estimation method sufficiently removes residual confounding, we introduce a nonparametric goodness-of-fit test based on the empirical distribution of p-values from NCOs. The diagnostic statistic, termed UDA, quantifies the L_1 distance between the empirical quantiles of observed p-values and the theoretical quantiles of the Uniform$(0, 1)$ distribution.

For a given estimation method, if it were unbiased, regular, and asymptotic linear, the standardized treatment effect estimate for each NCO should satisfy $\hat{\theta}_j/\hat{\sigma}_j \xrightarrow{d} \mathcal{N}(0, 1)$. Accordingly, we compute a two-sided p-value for each NCO as

$$p_j = 2\left\{1 - \Phi\left(\left|\frac{\hat{\theta}_j}{\hat{\sigma}_j}\right|\right)\right\}, \quad j = 1, \ldots, J,$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. These p-values are then used to assess the unbiasedness of the estimation method via uniformity-based diagnostics. Here, $\hat{\theta}_j$

is the point estimate produced by a given estimation procedure, and $\hat{\sigma}_j$ is the associated standard deviation. This framework applies not only to conventional estimators but also to the LOO-powered calibrated estimator introduced in our method.

Let $C(t)$ denote the empirical quantile function of the p-values, as plotted in the QQ plot. We define the uniformity deviation area as the area between the empirical QQ curve and the identity line $y = t$, given by the integral

$$\text{UDA} := \int_0^1 |C(t) - t| \, dt.$$

In practice, we approximate this area via a simple discretization. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(J)}$ denote the order statistics of the p-values. The UDA test statistic is computed as

$$T_J := \frac{1}{J} \sum_{j=1}^{J} \left| p_{(j)} - j/J \right|.$$

A large value of T_J suggests systematic deviation from uniformity and thus the estimation method is biased. In the *Supplementary Materials*, we compute the quantiles of the UDA statistic T_J under the null using a parametric bootstrap procedure. Here, we illustrate the procedure in the context of propensity score matching.

First, we follow the standard workflow to estimate the propensity score, perform matching, and compute the treatment effect estimates along with the resulting UDA statistic T_J. Second, for each bootstrap iteration indexed by b, $b = 1, \ldots, B$, we generate a new treatment vector $\{A_i^{(b)}\}_{i=1}^n$ by sampling from Bernoulli$(\hat{\pi}_i)$, where $\hat{\pi}_i$ denotes the estimated propensity score for unit i. Finally, using the resampled treatment assignments $\{A_i^{(b)}\}_{i=1}^n$, we re-estimate the propensity score, apply the matching procedure, and recompute $T_J^{(b)}$. Because the resampled treatment assignments depend only on the observed covariates, the resulting estimates are not subject to residual confounding bias. Repeating this procedure for B iterations yields an empirical distribution of $\{T_J^{(b)}\}_{b=1}^B$ that approximates the distribution of T_J under the null. This process enables us to compute critical values for T_J without relying on large-sample approximations.

Distributional set identification result

Under Assumption 1, the central $(1 - \alpha)$ quantile interval of Q directly yields a confidence region for the residual bias in the target estimand. That is, denoting by $q_{\alpha/2}$ and $q_{1-\alpha/2}$ the $\alpha/2$ and $1 - \alpha/2$ quantiles of Q, then

$$\mathbb{P}\left(q_{\alpha/2} \leq \theta_0^\star - \beta_A \leq q_{1-\alpha/2}\right) = 1 - \alpha,$$

implying that β_A is contained within $C_\alpha = [\theta_0^\star - q_{1-\alpha/2},\ \theta_0^\star - q_{\alpha/2}]$ with confidence level $1 - \alpha$. We now state a formal condition under which confidence intervals for the target causal effect β_A are identified via the distributional calibration framework.

Proposition 1. *Suppose that Assumption 1 holds. Then a* $(1-\alpha)$ *confidence interval* $C_\alpha = [\theta_0^\star - q_{1-\alpha/2},\ \theta_0^\star - q_{\alpha/2}]$ *for* β_A *can be identified such that* $\mathbb{P}(\beta_A \in C_\alpha) = 1 - \alpha$.

It is noteworthy that this form of identification departs from classical frameworks in causal inference, for example, point identification, where the target estimand β_A is determined exactly under strong assumptions, including a correctly specified DAG for the causal relationships among variables. Our approach also differs from set identification strategies [20], where a deterministic range of the estimand is obtained; for example, those used in sensitivity analysis and instrumental variable settings [21, 22].

By contrast, our framework yields a distributionally calibrated confidence region that reflects uncertainty in residual bias across multiple NCOs. We refer to this as *distributional set identification*, as it constructs a confidence region for the target estimand based on an estimated bias distribution. Unlike point identification

or sensitivity analysis, this approach provides an interval estimate with a coverage guarantee on the true causal effect based on the empirical data. Notably, when the residual bias distribution contracts to a point mass, the calibrated estimator becomes point-identified in the asymptotic limit; see the discussion following Theorem 1 for details.

In practice, we estimate $\theta_0^\star$ by $\hat{\theta}_0$, and infer quantiles of the distribution Q from the NCO estimates $\{\hat{\theta}_j\}_{j=1}^J$. In our implementation, we focus on the normal case $Q = \mathcal{N}(b, \tau^2)$, for which the parameters (b, τ^2) can be consistently estimated using method-of-moments estimators (see (2)). Nevertheless, the framework is not limited to the Gaussian case. More generally, the distribution Q can be estimated nonparametrically using techniques from empirical Bayes, such as the nonparametric maximum likelihood estimator (NPMLE) for a mixing distribution [23, 24]. These approaches allow inference on the quantiles of Q without imposing strong parametric assumptions, and can be readily incorporated into our calibration procedure.

Distributional calibration algorithm

For clarity, we separate the workflow into three transparent steps, which can be implemented with a few lines of code.

Step 1: Adjustment for observed confounders. Given an estimation method of β_A to adjust for observed confounders, such as outcome regression, inverse probability weighting/matching, or augmented inverse probability weighting, apply it to the outcome of primary interest and NCOs in the study cohort. Obtain the estimate $\hat{\theta}_j$ and its standard deviation $\hat{\sigma}_j$ for the outcome ($j = 0$) and each NCO ($j = 1, \ldots, J$).

Step 2: Estimation for unobserved confounder biases. Assume normal distribution $Q = \mathcal{N}(b, \tau^2)$ for the estimation biases, the method-of-moments estimators of (b, τ^2) are

$$\begin{aligned} \hat{\tau}^2 &= \max\left\{0, \frac{\sum_{j=1}^J \hat{\sigma}_j^{-2}(\hat{\theta}_j - \hat{\mu})^2 - (J-1)}{\sum_{j=1}^J \hat{\sigma}_j^{-2} - \sum_{j=1}^J \hat{\sigma}_j^{-4} / \sum_{j=1}^J \hat{\sigma}_j^{-2}}\right\}, \\ \hat{b} &= \frac{1}{\sum_{j=1}^J (\hat{\sigma}_j^2 + \hat{\tau}^2)^{-1}} \sum_{j=1}^J \frac{\hat{\theta}_j}{\hat{\sigma}_j^2 + \hat{\tau}^2}, \end{aligned} \tag{2}$$

where $\hat{\mu} = \left(\sum_{j=1}^J \hat{\sigma}_j^{-2}\right)^{-1} \sum_{j=1}^J \hat{\sigma}_j^{-2} \hat{\theta}_j$ denotes the fixed-effect meta estimator for b. Here, $\hat{b}$ can be viewed as the marginal mean of the bias effect.

Step 3: Distributional calibration. The calibrated estimate of β_A and its variance for the outcome of interest are

$$\hat{\theta}_0^{\text{cal}} \;=\; \hat{\theta}_0 - \hat{b} \;\text{ and }\; \widehat{\text{Var}}(\hat{\theta}_0^{\text{cal}}) \;=\; \hat{\sigma}_0^2 + \hat{\tau}^2,$$

yielding a $100(1-\alpha)\%$ confidence interval $\widehat{C}_\alpha = \hat{\theta}_0^{\text{cal}} \;\pm\; z_{1-\alpha/2}\sqrt{\hat{\sigma}_0^2 + \hat{\tau}^2}$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution.

Theoretical guarantees of DC

The following results show that the estimate of the bias distribution Q is consistent, and the confidence interval $\widehat{C}_\alpha$ achieves asymptotic coverage. Those results imply a valid inference of the target causal effect β_A by the proposed DC approach in the presence of unmeasured confounders.

Theorem 1 (Asymptotic coverage)**.** *Suppose that Assumption 1 holds with $Q = \mathcal{N}(b, \tau^2)$. If the regularity condition S.1 in the* Supplementary Materials *holds, then*

$$\frac{\hat{\theta}_0^{cal} - \beta_A}{\sqrt{\hat{\sigma}_0^2 + \hat{\tau}^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n, J \to \infty$. Consequently, the Wald-type confidence interval $\widehat{C}_\alpha$ achieves asymptotic $(1-\alpha)$ coverage for β_A.

As discussed in Proposition 1, one can only identify a valid $(1-\alpha)$ confidence interval for β_A; in this case, the calibrated point estimate $\hat{\theta}_0^{\text{cal}}$ is not meaningful as a consistent estimator. However, when the residual bias variance diminishes asymptotically, e.g., $\tau \to 0$, the proof of Proposition S.2 in the *Supplementary Materials* also implies that $\hat{b} - b \xrightarrow{p} 0, \hat{\tau}^2 - \tau^2 \xrightarrow{p} 0$. In this regime, the identified confidence interval contracts to a point as the sample size grows, and the calibrated estimate $\hat{\theta}_0^{\text{cal}}$ becomes a consistent estimator for β_A.

These results provide theoretical justification for the DC procedure as a valid bias correction layer. Notably, they demonstrate that the utility of DC does not hinge on full identifiability of the causal effect or the complete removal of confounding. Rather, it is sufficient to consistently estimate the distribution of residual bias across a collection of NCOs, enabling valid inference under partial confounding adjustment.

DC is agnostic to estimation methods and benefits from stronger covariate adjustment

In this section, we investigate two key problem of DC using linear models:

(i) Under what structural assumptions on the data generating mechanism do the biases of the outcome of interest and NCOs follow a common distribution, given a set of covariates and a pre-specified estimation method?

(ii) How does the level of covariate adjustment influence the efficiency and statistical power of the calibrated estimator?

To answer the two questions, we consider the following linear data-generating model:

$$Y_0(a) = \beta_{0,0} + \beta_A I(A = a) + \sum_{k=1}^{K} \gamma_{0,k} U_k + \gamma_{0,0}\varepsilon_0, \tag{3}$$

$$Y_j(1) = Y_j(0) = \beta_{j,0} + \sum_{k=1}^{K} \gamma_{j,k} U_k + \gamma_{j,0}\varepsilon_j, \tag{4}$$

where $U_{1:K} := (U_1, \ldots, U_K)$ denote K unmeasured confounders, and ε_j are independent additive noise terms for each outcome. The binary treatment $A \in \{0,1\}$ depends on $U_1, \ldots, U_K$, and the noise terms ε_j are independent of $(A, U_{1:K})$, with $\varepsilon_j \sim \mathcal{N}(0,1)$. The observed outcome is defined as $Y_0 = AY_0(1) + (1-A)Y_0(0)$ and $Y_j = AY_j(1) + (1-A)Y_j(0) = Y_j(1) = Y_j(0)$. Let P_j be the joint distribution of $(A, Y_j, U_{1:K})$ and $P_{j,t}$ the joint distribution of $(A, Y_j, U_{1:t})$.

To better understand the role of covariate adjustment in distributional calibration, we consider a setting in which the first t confounders, denoted by $U_{1:t}$, are assumed to be observed. That is, we observe n independent observations $(Y_{i,0}, Y_{i,1}, \ldots, Y_{i,J}, A_i, U_{i,1:t})$ from the model in (3) and (4) for $i = 1, \ldots, n$. This setup allows us to explore how adjusting varying levels of observed confounders influences both the distribution of estimation bias and the efficiency of the calibrated estimator.

Estimation agnosticism of DC. We argue that the validity of the proposed DC procedure does not depend on the specific choice of the estimation method. To elucidate the estimation-agnostic nature of DC, we consider a general class of estimation procedures. Let $\hat{\theta}_{j,t}$ denote the point estimate of the treatment effect for outcome j, obtained by applying a chosen estimation procedure to the observed data $\{(A_i, Y_{i,j}, U_{i,1:t})\}_{i=1}^n$. Define $\theta^\star_{j,t}$ as the probability limit of $\hat{\theta}_{j,t}$ as $n \to \infty$. It should be noted that some restrictions of estimation procedures should be made: procedures that, for example, arbitrarily assigning values for the estimates of NCOs would violate Assumption 1 and are excluded from consideration.

In this paper, we assume that the asymptotic limit of the treatment effect estimator for the outcome of interest takes the form

$$\theta^\star_{0,t} = \mathbb{E}\{Y_0(1)\} - \mathbb{E}\{Y_0(0)\} + \mathbb{E}\{M(\tilde{Y}_0, A, U_{1:t})\}, \tag{5}$$

where $\tilde{Y}_0 = A\tilde{Y}_0(1) + (1-A)\tilde{Y}_0(0)$ with $\tilde{Y}_0(1) = Y_0(1) - \mathbb{E}\{Y_0(1)\}$ and $\tilde{Y}_0(0) = Y_0(0) - \mathbb{E}\{Y_0(0)\}$, and $M(\cdot)$ represents some function uniquely determined by the estimation procedure. (5) holds if the estimating procedure can produce asymptotically unbiased estimates for the causal estimand when there are no unmeasured confounders.

For the observed NCOs, the causal effect of treatment is zero by design. Thus, the corresponding asymptotic limit takes the form

$$\theta^\star_{j,t} = \mathbb{E}\{M(\tilde{Y}_j, A, U_{1:t})\}, \tag{6}$$

where $\tilde{Y}_j = Y_j - \mathbb{E}(Y_j)$ denotes the centered version of the jth NCO. This formulation in (5) enables direct comparison of residual biases across outcomes on a common, mean-zero scale. Moreover, a variety of estimation procedures satisfy satisfy (5) and (6); for example, we show in the *Supplementary Materials* that the inverse propensity score weighting estimation and the outcome regression estimation meet this assumption.

Theorem 2. *Suppose the data are generated from the linear model specified in equations* (3)–(4)*, and that the first* t *confounders* $U_{1:t}$ *are observed. Assume the following conditions hold:*

(i) Confounders are mean zero, i.e., $\mathbb{E}(U_k) = 0$ *for* $k = 1, \ldots, K$*;*

(ii) (5) *and* (6) *hold for some unknown function* $M(\cdot)$ *uniquely determined by the estimation procedure;*

(iii) The coefficients $(\gamma_{j,0}, \ldots, \gamma_{j,K})^\top$ *are independently drawn across outcomes* j *from a shared distribution.*

Then, the limiting quantities $\theta^\star_{1,t}, \ldots, \theta^\star_{J,t}$ *and* $\theta^\star_{0,t} - \beta_A$ *are exchangeable and follow a common distribution determined by the law of* $(\gamma_{j,0}, \ldots, \gamma_{j,K})^\top$*.*

This result formalizes the key assumption underlying the proposed DC framework: namely, after adjusting for the observed confounders $U_{1:t}$, the residual biases across outcomes, quantified by the limiting estimands $\theta^\star_{j,t}$, are exchangeable and follow a shared distribution. Importantly, this property arises directly from the structure of the data-generating model instead of the specific estimation method.

Efficiency gain from stronger covariate adjustment. In what follows, we provide two examples, the propensity score weighting estimator and the outcome regression estimator, to illustrate that stronger covariate adjustment before DC can improve efficiency. Specifically, under inverse probability weighting (IPW), the estimated treatment effect for the jth outcome is

$$\hat{\theta}^{\text{ipw}}_{j,t} = \frac{1}{n}\sum_{i=1}^n \frac{Y_{i,j} I(A_i = 1)}{\hat{\pi}_t(U_{i,1:t})} - \frac{1}{n}\sum_{i=1}^n \frac{Y_{i,j} I(A_i = 0)}{1 - \hat{\pi}_t(U_{i,1:t})},$$

where $\hat{\pi}_t(U_{1:t})$ is an estimator for the propensity score $\pi_t(U_{1:t}) = \mathbb{P}(A = 1 \mid U_{1:t})$, and denote by $\pi_t^\star(U_{1:t})$ the probabilistic limit of $\hat{\pi}_t(U_{1:t})$. Then,

$$\hat{\theta}_{j,t}^{\text{ipw}} \xrightarrow{p} \theta_{j,t}^{\star\text{ipw}} = \mathbb{E}\Big\{ \frac{Y_j I(A=1)}{\pi_t^\star(U_{1:t})} - \frac{Y_j I(A=0)}{1-\pi_t^\star(U_{1:t})} \Big\}.$$

Alternatively, using outcome regression, the estimated treatment effect is

$$\hat{\theta}_{j,t}^{\text{or}} = \frac{1}{n}\sum_{i=1}^{n} \{\hat{m}_{j,t}(1, U_{i,1:t}) - \hat{m}_{j,t}(0, U_{i,1:t})\},$$

where $\hat{m}_{j,t}$ is an estimator for $m_{j,t}(a, U_{1:t}) = \mathbb{E}(Y_j \mid A = a, U_{1:t})$, and denote by $m_{j,t}^\star$ the probabilistic limit of $\hat{m}_{j,t}$. Then,

$$\hat{\theta}_{j,t}^{\text{or}} \xrightarrow{p} \theta_j^{\star\text{or}} = \mathbb{E}\{m_{j,t}^\star(1, U_{1:t}) - m_{j,t}^\star(0, U_{1:t})\}.$$

Next, we will show that when working models are correctly specified, the variability of the resultant asymptotic limit decreases as the adjustment level become stronger. Before stating the result, we define $\delta_k(u_{1:t}) = \mathbb{E}(U_k \mid A = 1, U_{1:t} = u_{1:t}) - \mathbb{E}(U_k \mid A = 0, U_{1:t} = u_{1:t})$.

Proposition 2. *Suppose that the data generating model* (3) *and* (4) *holds. If* $\pi_t = \pi_t^\star$ *and* $m_{j,t} = m_{j,t}^\star$*, then* (5) *and* (6) *hold with*

$$\theta_{j,t}^{\star ipw} = \theta_{j,t}^{\star or} = \sum_{k=t+1}^{K} \gamma_{j,k} \mathbb{E}\{\delta_k(U_{1:t})\}, \quad j = 1, \ldots, J,$$

$$\theta_{0,t}^{\star ipw} = \theta_{0,t}^{\star or} = \beta_A + \sum_{k=t+1}^{K} \gamma_{0,k} \mathbb{E}\{\delta_k(U_{1:t})\}.$$

These structured forms in Proposition 2 provide two insights. First, as shown in Proposition S.2 of the *Supplementary Material*, when the number of unmeasured confounders K is large, the assumption that each $\gamma_{j,k}$ follows the same distribution across outcomes j can be relaxed. In this high-dimensional regime, it suffices that the *weighted sums* of the means and variances of $\gamma_{j,k}$ converge to common constants across j. Under this condition, the residual biases remain approximately exchangeable, thereby justifying the use of distributional calibration. Second, as the number of observed confounders t increases, the number of unadjusted terms contributing to the residual bias decreases. This implies a reduction in the variability of the bias across outcomes, meaning that the uncertainty introduced by calibration diminishes. This property is naturally captured by the NCO calibration framework, which reflects the efficiency gain from stronger covariate adjustment.

Data availability

This study used population-level aggregate and de-identified data generated by the TriNetX platform. Due to data privacy, patient-level data cannot be shared.

Code availability

To facilitate adoption, we provide an open-source R package, *debiasedTrialEmulation*, available at `https://cran.r-project.org/web/packages/debiasedTrialEmulation/index.html`. While originally developed to support target trial emulation workflows, the package is fully modular: the distributional calibration module can be used independently of the trial emulation components. This allows users to

incorporate our bias correction method into a wide range of causal inference pipelines. The package includes functions for fitting standard adjustment methods, estimating bias distributions from NCOs, and generating calibrated confidence intervals. Comprehensive documentation and example workflows are provided to support integration into large-scale real-world evidence studies.

Funding

This work was supported in part by National Institutes of Health (U01TR003709, U24MH136069, U24AG098157, RF1AG077820, R01AG073435, R01LM013519, R01DK128237, R21AI167418, R21EY034179).

References

[1] Steven P Marso, Stephen C Bain, Agostino Consoli, Freddy G Eliaschewitz, Esteban Jódar, Lawrence A Leiter, Ildiko Lingvay, Julio Rosenstock, Jochen Seufert, Mark L Warren, et al. Semaglutide and cardiovascular outcomes in patients with type-2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844, 2016.

[2] Reimar W Thomsen, Aurelie Mailhac, Julie B Løhde, and Anton Pottegård. Real-world evidence on the utilization, clinical and comparative effectiveness, and adverse effects of newer GLP-1RA-based weight-loss therapies. *Diabetes, Obesity and Metabolism*, 27:66–88, 2025.

[3] Truveta. GLP-1 RA Prescription Trends, September 2025, 2025.

[4] Mohit Sodhi, Ramin Rezaeianzadeh, Abbas Kezouh, and Mahyar Etminan. Risk of gastrointestinal adverse events associated with glucagon-like peptide-1 receptor agonists for weight loss. *JAMA*, 330(18):1795–1797, 2023.

[5] William Wang, Nora D Volkow, Nathan A Berger, Pamela B Davis, David C Kaelber, and Rong Xu. Association of semaglutide with risk of suicidal ideation in a real-world cohort. *Nature Medicine*, 30(1):168–176, 2024.

[6] Huilin Tang, Ying Lu, William T Donahoo, Sarah C Westen, Yong Chen, Jiang Bian, and Jingchuan Guo. Glucagon-like peptide-1 receptor agonists and risk for depression in older adults with type 2 diabetes: A target trial emulation study. *Annals of Internal Medicine*, 178(3):315–326, 2025.

[7] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

[8] Benjamin F Arnold and Ayse Ercumen. Negative control outcomes: A tool to detect bias in randomized trials. *JAMA*, 316(24):2597–2598, 2016.

[9] Paul R Rosenbaum. The role of known effects in observational studies. *Biometrics*, 45(2):557–569, 1989.

[10] Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B Owen. Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics*, 45(5):1863, 2017.

[11] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[12] Wang Miao, Wenjie Hu, Elizabeth L Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023.

[13] Lauren E Dang, Susan Gruber, Hana Lee, Issa J Dahabreh, Elizabeth A Stuart, Brian D Williamson, Richard Wyss, Iván Díaz, Debashis Ghosh, Emre Kıcıman, et al. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, 7(1):e212, 2023.

[14] Rohini K Hernandez, Cathy W Critchlow, Nancy Dreyer, Timothy L Lash, Robert F Reynolds, Henrik T Sørensen, Jeff L Lange, Nicolle M Gatto, Rachel E Sobel, Edward Chia-Cheng Lai, et al. Advancing principled pharmacoepidemiologic research to support regulatory and healthcare decision making: the era of real-world evidence. *Clinical Pharmacology & Therapeutics*, 117(4):927–937, 2025.

[15] Erica A Voss, Richard D Boyce, Patrick B Ryan, Johan van der Lei, Peter R Rijnbeek, and Martijn J Schuemie. Accuracy of an automated knowledge base for identifying drug adverse reactions. *Journal of Biomedical Informatics*, 66:72–81, 2017.

[16] Rohan Khera, Arya Aminorroaya, Lovedeep Singh Dhingra, Phyllis M Thangaraj, Aline Pedroso Camargos, Fan Bu, Xiyu Ding, Akihiko Nishimura, Tara V Anand, Faaizah Arshad, et al. Comparative effectiveness of second-line antihyperglycemic agents for cardiovascular outcomes: A multinational, federated analysis of LEGEND-T2DM. *Journal of the American College of Cardiology*, 84(10):904–917, 2024.

[17] Guangyong Zou. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7):702–706, 2004.

[18] Sander Greenland. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, 125(5):761–768, 1987.

[19] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

[20] Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[21] Jesse Y Hsu, Dylan S Small, and Paul R Rosenbaum. Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108(501):135–148, 2013.

[22] Dylan S Small, Zhiqiang Tan, Roland R Ramsahai, Scott A Lorch, and M Alan Brookhart. Instrumental variable estimation with a dtochastic monotonicity assumption. *Statistical Science*, 32(4):561–579, 2017.

[23] Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

[24] Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87 (1):1–32, 2025.