

Transport G-Computation: A Distributional Approach to Longitudinal Causal Inference via Optimal Transport

Yuanyuan Huang

Jiangyin People's Hospital

Tianpu Feng

Jiangsu Ocean University

Jue Zhang

Kangda College of Nanjing Medical University

Xiaoxue Song

Jiangsu Ocean University

Xijun He

hy12res@163.com

Jiangyin People's Hospital

Research Article

Keywords: Transport g-computation, longitudinal causal inference, observational data, optimal transport, Wasserstein distance, simulation study

Posted Date: July 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10203429/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Transport G-Computation: A Distributional Approach to Longitudinal Causal Inference via Optimal Transport

Yuanyuan Huang¹, Tianpu Feng², Jue Zhang³, Xiaoxue Song⁴, Xijun He^{1*}

¹ Department of Emergency Medicine, Jiangyin People's Hospital, Jiangyin, Jiangsu, China.

² Jiangsu Ocean University, Lianyungang, Jiangsu, China.

³ Kangda College of Nanjing Medical University, Lianyungang 222000, Jiangsu, China.

⁴ Jiangsu Key Laboratory of Marine Pharmaceutical Compound Screening, College of Pharmacy, Jiangsu Ocean University, Lianyungang 222005, Jiangsu, China.

*Correspondence: Xijun He, Department of Emergency Medicine, Jiangyin People's Hospital, Jiangyin, Jiangsu, China. Email: hy12res@163.com

Abstract

Background: Longitudinal causal inference using observational health data is complicated by time-dependent confounding, feedback between treatment assignment and evolving patient states, and heterogeneous post-intervention outcome distributions. Conventional g-computation usually focuses on mean-level effects and may miss clinically relevant distributional changes. We propose Transport G-Computation (TGC), a distributional framework for estimating Wasserstein causal effects between terminal potential outcome distributions. **Methods:** TGC combines longitudinal g-computation with local optimal transport. At each time point and

treatment level, local nearest-neighbor samples are used to construct empirical transition libraries, and entropy-regularized optimal transport is applied to approximate conditional transition distributions. Intervention trajectories are recursively simulated under static treatment regimes, and the Wasserstein distance between terminal empirical distributions is used as the causal estimand. TGC was evaluated in four simulation scenarios and compared with parametric g-computation using bias, root mean square error (RMSE), relative bias, runtime, and failure rate across 100 replications. **Results:** Parametric g-computation performed best in the linear Gaussian and Gaussian mixture scenarios, with RMSE values of 0.075 and 0.080, compared with 0.659 and 0.252 for TGC. TGC showed its strongest advantage in the confounded feedback scenario, reducing RMSE from 1.821 to 0.486 and absolute bias from 1.813 to 0.468. In the nonlinear heteroscedastic scenario, both methods showed substantial error. All replications completed successfully, but TGC required substantially longer runtime. **Conclusions:** TGC offers a flexible distributional approach for longitudinal causal inference and may be useful when feedback confounding and model misspecification limit parametric g-computation. Further work is needed to improve tuning, efficiency, and practical implementation for observational medical research.

Keywords: Transport g-computation; longitudinal causal inference; observational data; optimal transport; Wasserstein distance; simulation study

Background

Longitudinal causal inference is among the most common and challenging problems in observational studies^{1,2}. Unlike cross-sectional settings, treatment assignment, state evolution, and future outcomes in longitudinal data are typically governed by complex dynamic dependencies: current treatments influence subsequent states and may, through feedback from states, continue to affect future treatments, thereby inducing time-varying confounding^{3,4}.

Classical longitudinal g-computation provides a foundational framework for identification and estimation in such settings^{1,2}. However, its practical performance is highly dependent on the correct specification of the conditional transition model⁵. When the true data-generating mechanism exhibits nonlinearity, multimodality, heteroscedasticity, or complex feedback structures, parametric models often struggle to adequately capture the dynamic transition process^{6,7}, leading to systematic bias in the estimation of terminal causal effects.

On the other hand, most existing longitudinal causal methods focus primarily on mean-type effects, such as the average treatment effect, expected outcome differences, or changes in regression coefficients. Yet in many practical applications, the impact of an intervention is not restricted to the mean^{8,9}; it may also manifest in the shape of the outcome distribution, such as distributional shifts, changes in dispersion, or the reshaping of multimodal structures. In such cases, comparing means alone is often insufficient to address the research question. Causal effects at the distributional level thus constitute a more natural and informative target^{8,10}. The Wasserstein distance, as a geometric tool for measuring discrepancies between two distributions, simultaneously reflects location, spread, and shape changes^{10,11}, making it well suited as an overall effect measure for post-intervention distributional changes in longitudinal settings.

Despite its favorable interpretability for distributional comparisons, systematically incorporating the Wasserstein distance into longitudinal causal inference presents several challenges. First, the terminal distribution under intervention must be obtained recursively through the longitudinal transition mechanism, which requires estimating not only the terminal distribution itself but also the conditional transition kernel at each step. Second, if a global parametric model is adopted, the method's performance will be heavily constrained by the correctness of the model specification⁵. However, if a fully nonparametric approach is employed, issues such as high dimensionality,

data sparsity, and positivity violations arise^{6,12}. Moreover, under the recursive temporal structure, early estimation errors propagate backward and accumulate at the terminal time, imposing more stringent demands on the stability of conditional transition estimation¹³.

To address these issues, we propose a novel distributional-level longitudinal causal estimation framework: Transport G-Computation (TGC). This method combines the idea of optimal transport with longitudinal g-computation^{14,15}. At each time point and for each treatment arm, it constructs a nonparametric approximation of the conditional transition distribution through local neighborhoods and OT couplings^{15,16}. Compared with traditional parametric g-computation, TGC does not rely on global linear models; instead, it preserves the multimodality and stochastic structure of the empirical distribution through localization and soft matching¹¹. Compared with simple nearest-neighbor or kernel regression, TGC leverages optimal transport to establish richer local mappings between source and target distributions^{14,17}, making it better suited for complex longitudinal distributional transitions.

The main contributions of this paper are as follows. First, we introduce the Wasserstein distance into the definition of longitudinal causal effects, constructing an estimation target that directly addresses terminal distributional changes. Second, we propose the TGC method, which combines local kernel smoothing with optimal transport coupling, to estimate the terminal distribution under static intervention. Third, we provide corresponding identification and theoretical analysis, showing that the target is identifiable under standard assumptions, and we discuss the properties of local conditional transition estimation and recursive error propagation. Fourth, we compare the performance of TGC and parametric g-computation across multiple simulation scenarios; the results indicate that TGC is more advantageous in dynamic feedback confounding scenarios, whereas it is not necessarily superior to traditional methods in simple settings with correctly

specified models. Overall, this paper aims to provide an estimation framework for distributional-level longitudinal causal inference that is simultaneously interpretable, flexible, and amenable to diagnostics.

Methods

2.1 Problem Setup and Target Quantity

We consider a discrete-time longitudinal causal inference problem^{1,2}. For each individual, at time points $t = 0, \dots, T - 1$, we observe a state vector $L_t \in \mathbb{R}^d$ and a binary treatment $A_t \in \{0, 1\}$.

Denote the potential outcomes by $L_t^{(a)}$, representing the longitudinal trajectory under the static intervention $A_t \equiv a$. Rather than traditional mean-type causal effects^{8,9}, the focus of this paper is on the distributional-level difference at the terminal time⁸, namely comparing the distributions of the terminal potential outcomes $L_T^{(1)}$ and $L_T^{(0)}$ under the two static interventions $a = 1$ and $a = 0$.

To this end, we define the distributional-level causal effect as the second-order Wasserstein distance^{10,11}:

$$\text{WCE} = W_2(\mathbb{P}(L_T^{(1)}), \mathbb{P}(L_T^{(0)}))$$

This quantity not only reflects shifts in distributional location but also captures changes in dispersion and multimodal structure, making it appropriate for longitudinal processes with strong heterogeneity, nonlinear feedback, and non-Gaussian errors. Compared with approaches that only compare means or regression coefficients, the WCE more directly addresses the question of whether and to what extent the overall distribution changes after intervention.

2.2 Identification Assumptions

Under the following standard causal assumptions², the terminal distribution under static intervention can be identified from observed data:

- Consistency: If an individual's actually received treatment path coincides with the intervention regime, the observed outcome equals the corresponding potential outcome;
- Sequential ignorability: Given the history of states L_t , the current treatment A_t is independent of future potential outcomes;
- Positivity: Within the support of observable states, every treatment level has non-zero probability¹⁸.

Under these conditions, the terminal distribution for the static intervention $A_t \equiv a$ satisfies the longitudinal g-formula^{1,2}:

$$\mathbb{P}(L_T^{(a)}) = \int \mathbb{P}(L_0) \prod_{t=0}^{T-1} \mathbb{P}(L_{t+1} \mid L_t, A_t = a) dL_{0:T-1}$$

Thus, the key to estimating the WCE lies in the reliable estimation of the conditional transition kernel $\mathbb{P}(L_{t+1} \mid L_t, A_t = a)$; once this transition kernel can be consistently estimated, one can obtain the terminal distribution under intervention via forward simulation and subsequently compute the Wasserstein distance.

2.3 Transport G-Computation

Building upon the above identification result, we propose Transport G-Computation (TGC). The core idea of TGC is to avoid imposing strong parametric assumptions on the longitudinal transition mechanism¹⁹, and instead, at each time point and for each treatment arm, construct a

129 nonparametric approximation of the conditional transition distribution through a localized
 130 optimal transport mechanism. More specifically, for a given time point t and treatment level a ,
 131 we extract from the observed data all sample pairs satisfying $A_t = a$:

$$132 \quad D_{t,a} = \left\{ \left(L_t^{(i)}, L_{t+1}^{(i)} \right) \mid A_t^{(i)} = a \right\},$$

133 and treat them as an empirical transition library between source states and target states. If the
 134 sample size for a particular treatment arm at a particular time point is too small, the region is
 135 deemed to suffer from positivity violation²⁰, and the algorithm does not attempt forced
 136 extrapolation; instead, it explicitly reports failure to avoid spuriously precise estimates in non-
 137 identifiable regions.

138 For any query state x , TGC first identifies local nearest neighbors in the source state set¹⁹,
 139 thereby forming a local subsample around x . The purpose of this step is to restrict estimation to
 140 the observed region near the current state, reducing the bias induced by global modeling in high
 141 dimensions. Subsequently, an entropically regularized optimal transport problem is solved on
 142 this local subsample^{14,21}. Let the local source samples be $\{x_j\}_{j=1}^n$ and the local target samples be
 143 $\{y_j\}_{j=1}^n$; by solving

$$144 \quad P^* = \arg \min_{P \in \Pi(a,b)} \langle P, M \rangle + \lambda \text{KL}(P \parallel ab^\top)$$

145 we obtain a soft-matching coupling from source to target, where M is the squared Euclidean cost
 146 matrix, λ is the entropic regularization strength, and $\Pi(a, b)$ denotes the set of couplings
 147 satisfying marginal constraints. This coupling matrix, on the one hand, preserves the
 148 multimodality of the local transition distribution, and on the other hand, enhances numerical
 149 stability through entropic regularization^{14,21}.

150 To apply this coupling to a specific query state x , TGC further constructs kernel weights with
 151 respect to the local source states. We employ Gaussian kernel weights¹⁹:

$$152 \quad w_j(x) \propto \exp\left(-\frac{|x - x_j|^2}{2h^2}\right)$$

153 where the bandwidth h^2 is adaptively determined by the median of local distances. The OT
 154 coupling is then row-normalized and combined with the kernel weights to form a local
 155 conditional distribution approximation over the target states¹⁵. Finally, the next-time state L_{t+1} is
 156 sampled from this weighted distribution. This mechanism can be understood as a combination of
 157 “local kernel smoothing + OT matching”: kernel weights handle localization, optimal transport
 158 preserves the structure of the empirical distribution, and stochastic sampling allows distributional
 159 uncertainty to propagate naturally through the longitudinal recursion¹⁵.

160 **2.4 Recursive Simulation of Intervention Trajectories**

161 Once the conditional transition estimates are obtained, TGC recursively generates the entire
 162 trajectory under static intervention in chronological order. Given an initial state sample L_0 , under
 163 the intervention $A_t \equiv a$, the algorithm applies the above local conditional transition estimator at
 164 each time point, yielding

$$165 \quad L_{t+1}^{(a)} \sim \widehat{\mathbb{P}}(\cdot | L_t^{(a)}, A_t = a), \quad t = 0, \dots, T - 1.$$

166 Repeating this procedure produces two sets of terminal samples $\{L_T^{(1)}\}$ and $\{L_T^{(0)}\}$. Finally, the
 167 second-order Wasserstein distance is computed based on the empirical distributions to obtain the
 168 $\widehat{\text{WCE}}$ ¹¹.

2.5 Comparison Method: Parametric G-Computation

To evaluate the performance of TGC, we also implement a parametric g-computation baseline³. This method fits linear transition models separately for each time point and each dimension, and performs forward simulation under static intervention to obtain the terminal distribution and its Wasserstein distance. This baseline represents the traditional parametric longitudinal causal inference strategy; its advantages are simplicity of implementation and clear interpretability, but its performance depends on whether the transition model is correctly specified⁸. When the true transition exhibits pronounced nonlinearity, multimodality, or heteroscedasticity, parametric models typically produce systematic bias. In contrast, TGC reduces reliance on the global model form through the local OT mechanism and is better suited to scenarios with complex distributional transitions.

2.6 Simulation Design and Evaluation Metrics

We construct multiple data-generating mechanisms to systematically evaluate the performance of different methods under various longitudinal transition structures, including linear Gaussian transitions, multimodal mixture transitions, treatment effects with feedback, and nonlinear heteroscedastic processes. Treatment assignment is driven by the current state, thereby creating a confounded treatment assignment mechanism to test the robustness of methods under observed confounding. Each scenario is repeated independently multiple times, and the following metrics are reported: estimation bias, absolute bias, RMSE, average runtime, and failure rate. If a method is unable to produce an estimate in a particular repetition due to positivity violation or numerical instability, that repetition is recorded as a failure and reported separately in the summary.

190 **Theoretical properties**

191 **3.1 Identifiability**

192 We first establish the identifiability of the WCE. Since the target estimand in this paper is
193 defined as the Wasserstein distance between the terminal potential outcome distributions under
194 static interventions, the WCE is identifiable as long as the terminal distributions themselves are
195 identifiable²².

196 **Assumption 1 (Consistency).**

197 If an individual's actually observed treatment path coincides with a static intervention a , then
198 their observed outcome equals the corresponding potential outcome.

199 **Assumption 2 (Sequential Ignorability).**

200 For each time point t , the current treatment A_t is independent of future potential outcomes
201 conditional on the history L_t .

202 **Assumption 3 (Positivity).**

203 Within the support of the states required by the study, every treatment level $a \in \{0,1\}$ has non-
204 zero conditional probability^{18,20}.

205 **Theorem 1 (Identifiability of the WCE).**

206 Under Assumptions 1–3, the terminal distribution $\mathbb{P}(L_T^{(a)})$ under static intervention $A_t \equiv a$ is
207 uniquely identified by the longitudinal g-formula, and hence

$$208 \quad \text{WCE} = W_2(\mathbb{P}(L_T^{(1)}), \mathbb{P}(L_T^{(0)}))$$

209 is also identifiable.

210 By consistency and sequential ignorability, the potential outcome distribution can be written as
211 an integral over conditional transition kernels, namely the longitudinal g-formula^{1,2}. Positivity
212 ensures that the conditional distributions in the integral can be estimated from the observed data
213 over the target support, so the terminal distribution under intervention is identifiable. Since the
214 Wasserstein distance is a deterministic function of two identifiable distributions, the WCE is also
215 identifiable^{10,22}.

216 **3.2 Approximation Properties of Local Transition Estimation**

217 The TGC estimates a nonparametric approximation to the local conditional transition kernel¹⁵.
218 Intuitively, as long as the true conditional transition is sufficiently smooth in the state variables
219 and there are enough samples within the local neighborhood, the estimator constructed from
220 nearest-neighbor truncation, kernel weighting, and OT coupling should approximate the true
221 conditional transition.

222 **Assumption 4 (Local Smoothness).**

223 For each time point t and treatment level a , the true conditional transition kernel
224 $\mathbb{P}(L_{t+1} \mid L_t = x, A_t = a)$ is locally Lipschitz continuous in x ¹⁷.

225 **Assumption 5 (Local Support Sufficiency).**

226 In the neighborhood of the target point x , the conditional density of the source state is bounded
227 below by a positive constant, and as the sample size increases, the covering radius of the local
228 nearest-neighbor set shrinks to zero²³.

Proposition 1 (Consistent Approximation of the Local Conditional Transition).

Under Assumptions 4–5, and provided that the number of nearest neighbors k grows slowly with the sample size and the kernel bandwidth shrinks at an appropriate rate, the local conditional transition distribution constructed by TGC converges to the true conditional transition kernel in the sense of weak convergence¹¹.

Nearest-neighbor truncation restricts estimation to the vicinity of the target point, and the global bias decreases as the neighborhood radius shrinks²⁴. Kernel weights suppress the influence of distant samples, so that the estimation is primarily governed by the local geometry¹⁹. The OT coupling establishes soft matching between local empirical distributions, preserving the multimodality and stochasticity of the target distribution^{17,25}. Under sufficient local sample size, the discrepancy between the empirical coupling and the true transition kernel can be controlled¹⁴.

3.3 Recursive Error Propagation

The terminal distribution of the TGC is not estimated in a single step, but is obtained through longitudinal recursive simulation. Therefore, the estimation error of the conditional transition at each time point propagates forward. This property is important because it explains why every step of the method must be as stable as possible, and why problems over long time horizons are more challenging.

Assumption 6 (Recursive Stability).

There exists a constant $C_0 > 0$ such that the true transition operator at each time point satisfies a local Lipschitz-type stability in the Wasserstein metric¹⁷.

249 **Proposition 2 (Recursive Control of the Terminal Error).**

250 Let ε_t denote the conditional transition estimation error at step t in the Wasserstein metric. Under
251 Assumption 6, there exists a constant $C > 0$ such that the terminal distribution error satisfies

252
$$W_2\left(P\left(L_T^{(a)}\right), \hat{P}\left(L_T^{(a)}\right)\right) \leq C \sum_{t=0}^{T-1} \varepsilon_t.$$

253 Consequently, if the conditional transition estimate at each step is consistent, then the terminal
254 distribution estimate is also consistent¹³.

255 The longitudinal recursion is viewed as a composition of distribution operators²⁶. By the local
256 Lipschitz stability in Assumption 6, the error at step t propagates to step $t + 1$ with a factor no
257 greater than C_0 ; meanwhile, the newly introduced local estimation error ε_t at each step
258 accumulates in an additive manner. Recursive application of the triangle inequality yields the
259 error bound stated above.

260 **3.4 Diagnosability Under Positivity Violation**

261 Because TGC explicitly relies on local sample support within each treatment arm, it does not
262 silently extrapolate in regions where samples are sparse. This is a methodological advantage, as
263 it prevents the situation where “results appear to exist, yet there is no identification basis”^{20,27}.

264 **Proposition 3 (Explicit Diagnosis of Positivity Violation).**

265 If the effective sample size in a treatment arm at a given time point falls below the minimum
266 support threshold, the conditional transition kernel over that region cannot be reliably estimated.
267 TGC explicitly flags such instances as a positivity violation, rather than resorting to implicit
268 estimation through pooled fallback or borrowing strength across arms.

When the sample size is insufficient, the local empirical distribution cannot provide adequate neighborhood coverage for the target point, and hence the stability and reproducibility of the local transition kernel estimate cannot be guaranteed. TGC makes this unidentifiable scenario explicit through a minimum sample threshold, thereby preventing spurious estimation conclusions.

3.5 Theoretical Explanation of Methodological Advantages

Synthesizing the properties above, the advantages of TGC can be summarized in three points. First, TGC does not rely on a global parametric model, but instead alleviates model misspecification bias through local kernel smoothing. Second, the OT coupling preserves the multimodality and stochasticity of the target conditional distribution, thereby offering greater expressiveness than point regression or nearest-neighbor imputation. Third, TGC explicitly exposes positivity violations, avoiding uncontrolled extrapolation in regions lacking identification support. Consequently, when the true transition mechanism exhibits nonlinear, multimodal, or heteroscedastic structures, TGC is expected to deliver more stable distribution-level estimates than parametric g-computation.

Results

We compared the performance of TGC and parametric g-computation (Param_gcomp) under four simulation scenarios, including linear Gaussian transitions, Gaussian mixture transitions, confounded scenarios with feedback, and nonlinear heteroscedastic scenarios. Each scenario was replicated 100 times, with evaluation metrics comprising the true WCE, mean estimate, average bias, mean absolute deviation, RMSE, relative bias, average runtime, and failure rate. Overall, parametric g-computation demonstrated greater stability in most scenarios, whereas TGC

showed a clear advantage only in the dynamic feedback scenario; meanwhile, the computational cost of TGC was substantially higher, yet all replications completed successfully without any increase in failure rate. Detailed numerical results for both methods across all scenarios are summarized in **Table 1**.

Table 1. Performance comparison between TGC and parametric g-computation across four simulation scenarios.

Scenario	Method	True WCE	Mean Est.	Bias	RMSE	Rel. Bias	Time (s)
a) Linear	Param_gcomp	2.896	2.914	0.018	0.075	0.006	1.1
Gaussian	TGC	2.896	2.241	-0.655	0.659	0.226	179.6
b) Gaussian	Param_gcomp	2.887	2.911	0.024	0.08	0.008	1.1
Mixture	TGC	2.887	2.649	-0.238	0.252	0.083	136.3
c) Confounded	Param_gcomp	9.478	11.291	1.813	1.821	0.191	1.1
Feedback	TGC	9.478	9.01	-0.468	0.486	0.049	234.7
d) Nonlinear	Param_gcomp	6.92	10.565	3.645	3.658	0.527	1.2
Hetero	TGC	6.92	3.277	-3.643	3.645	0.526	181.8

Note: All scenarios were run with 100 attempted replications, all of which completed successfully (failure rate = 0).

In the linear Gaussian scenario, both methods recovered the true WCE reasonably well. For Param_gcomp, the true WCE was 2.896, the mean estimate was 2.914, the average bias was merely 0.018, the mean absolute deviation was 0.057, and the RMSE was 0.075. By contrast, the mean estimate of TGC was 2.241, exhibiting a notably negative bias, with an average bias of -0.655, and the mean absolute deviation and RMSE were 0.655 and 0.659, respectively. This result indicates that when the true transition structure aligns well with a linear parametric model, traditional parametric g-computation is in fact preferable, whereas the local OT smoothing employed by TGC may introduce additional estimation bias in this simple scenario.

In the Gaussian mixture scenario, Param_gcomp maintained satisfactory accuracy, with an average bias of 0.024, a mean absolute deviation of 0.063, and an RMSE of 0.080. The mean

estimate of TGC was 2.649, corresponding to an average bias of -0.238, a mean absolute deviation of 0.238, and an RMSE of 0.252. Although TGC still approximated the true WCE reasonably well, its error was markedly higher than that of the parametric baseline. In other words, under the current simulation settings, TGC did not exhibit the anticipated advantage in multi-modal distribution scenarios, suggesting that its performance may be sensitive to neighborhood size, kernel bandwidth, or OT regularization parameters.

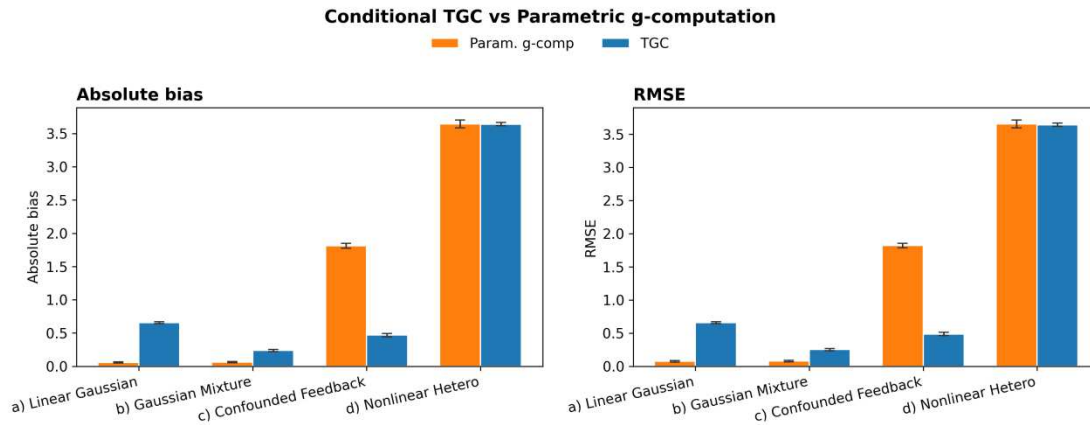
In the confounded feedback scenario, TGC demonstrated its most pronounced advantage. In this scenario, the true WCE was 9.478, and the mean estimate of Param_gcomp was 11.291, with the average bias reaching 1.813; the mean absolute deviation and RMSE were likewise 1.813 and 1.821. By comparison, the mean estimate of TGC was 9.010, with an average bias of -0.468, a mean absolute deviation of 0.468, and an RMSE of 0.486, substantially outperforming the parametric baseline. This result illustrates that when treatment assignment exhibits feedback dependence on historical states and the dynamic confounding structure is more complex, the local conditional transition estimation of TGC is more effective in reducing systematic bias arising from model misspecification.

In the nonlinear heteroscedastic scenario, both methods faced considerable modeling challenges. The mean estimate of Param_gcomp was 10.565, substantially overestimating the true WCE of 6.920, with an average bias of 3.645 and an RMSE of 3.658. The mean estimate of TGC was 3.277; although the direction of bias was opposite, the absolute bias was similarly large at 3.643, with an RMSE of 3.645. That is, in this scenario, TGC did not significantly outperform the parametric baseline; rather, both methods exhibited conspicuous systematic errors. This suggests that for strongly nonlinear, heteroscedastic settings with complex distributional structures, the

current version of TGC still requires further hyperparameter tuning or methodological refinement.

Synthesizing the estimation accuracy across the four scenarios, the performance of the two methods under different data-generating mechanisms exhibited marked differences (**Figure 1**). Figure 1 provides an intuitive visualization of the comparison of average bias and RMSE across scenarios, further corroborating that TGC offered significantly superior robustness relative to the parametric baseline only in the confounded feedback scenario, while encountering challenges of varying degrees in the other scenarios.

Figure 1. Comparison of bias and root mean square error (RMSE) between the transport g-computation (TGC) and parametric g-computation across four simulation scenarios.

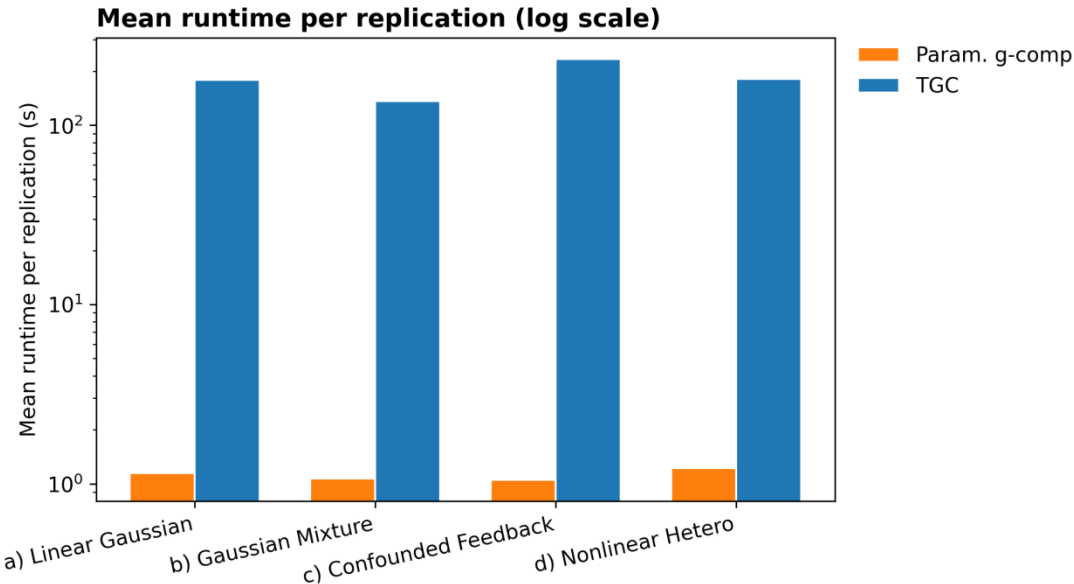


Note: This figure illustrates the average bias and RMSE of both methods under four distinct scenarios: linear Gaussian, Gaussian mixture, feedback confounding, and nonlinear heteroscedasticity. TGC demonstrated the smallest bias and error in Scenario c, whereas relatively larger errors were observed in Scenarios a and d.

From the perspective of computational efficiency, the runtime of TGC was markedly higher than that of Param_gcomp (**Figure 2**). As shown in **Figure 2**, the average runtime of Param_gcomp

was approximately 1.05–1.22 seconds per replication, whereas that of TGC ranged from approximately 136.35 to 234.69 seconds per replication. In other words, TGC was roughly two orders of magnitude slower under the current implementation. Although such overhead is acceptable in methodological research, further optimization of the implementation would be necessary for large-scale data applications or simulations with a high number of replications, particularly in nearest-neighbor search, local OT solving, and parallelization strategies.

Figure 2. Comparison of average runtime between TGC and Param_gcomp across different scenarios.



Note: The y-axis represents the average runtime (in seconds) on a logarithmic scale. TGC exhibited consistently and markedly higher runtime than Param_gcomp across all scenarios, indicating a substantial computational overhead.

It is worth noting that the number of successful replications was 100 for both methods across all scenarios, with a failure rate of 0, indicating that the numerical procedures were stable under the

current experimental settings, with no evident positivity collapse or numerical failure. The MCSE was also small overall, suggesting good stability of the replicated simulation results. Taken together, these results support a relatively conservative yet credible conclusion: TGC has advantages in complex longitudinal confounding scenarios with dynamic feedback, but does not uniformly outperform parametric g-computation across all complex distributional settings; at the same time, TGC entails substantially higher computational costs. Therefore, a more appropriate characterization in the paper is not “universally superior to the baseline,” but rather “more robust under feedback confounding structures, comparable to or slightly inferior to the baseline in other scenarios, and with considerable computational overhead.”

Discussion

The Transport G-Computation proposed in this paper aims to integrate optimal transport with longitudinal g-computation to estimate the Wasserstein causal effect between terminal distributions under static interventions^{10,22}. Compared with traditional parametric g-computation, the core advantage of TGC lies in its avoidance of reliance on global linear models; instead, it characterizes conditional transition distributions through local neighborhoods and OT couplings¹⁶, and is therefore theoretically better suited for estimating distributional causal effects under multi-modal, nonlinear, and complex feedback structures. Our simulation results generally support this conceptual premise, but also reveal that the method in its current implementation does not dominate in all scenarios, particularly in simpler settings or those well approximated by linear models, where the parametric baseline in fact performs better.

First, in the linear Gaussian scenario, parametric g-computation clearly outperformed TGC. This result is not surprising, because this scenario is highly consistent with the linear model

specification, allowing the parametric method to fully leverage the correct model structure and thereby achieve lower bias and smaller RMSE¹⁰. Conversely, because TGC incorporates steps such as local neighborhood selection, OT regularization, and stochastic sampling, it may introduce unnecessary smoothing and additional variance in simple scenarios, leading to estimates that deviate from the true value. This phenomenon serves as a reminder that TGC is not a “universally stronger” black-box replacement in all circumstances, but rather a nonparametric method more oriented toward complex data structures. When the model is correctly specified and the target distribution is relatively simple, traditional parametric methods remain highly competitive⁷.

Second, in the Gaussian mixture scenario, TGC still failed to demonstrate a clear advantage. Theoretically, such multi-modal structures represent precisely the settings in which optimal transport methods are most promising, because OT preserves distributional shape rather than merely mean information^{11,17}. Nevertheless, the empirical results show that the current version of TGC achieved only moderate performance in this scenario, failing to surpass the parametric baseline. One plausible explanation is that, although the true outcome distribution is multi-modal, the terminal WCE itself may not be sufficiently sensitive to this multi-modal structure, or the current choices of local neighborhood size, kernel bandwidth, and entropy regularization strength are not optimally matched to this data-generating mechanism¹⁶. In other words, the potential advantages of TGC may require more refined hyperparameter tuning, or may be more fully realized in higher-dimensional settings with more pronounced nonlinear distributional shifts.

Third, in the confounded feedback scenario, TGC exhibited the most stable and compelling advantage. This result is most consistent with the original design motivation of the method.

407 Feedback confounding implies that treatment assignment is influenced not only by the current
408 state but also by complex interactions with historical trajectories, which substantially increases
409 the risk of model misspecification for parametric models^{3,28}. Under a linear regression
410 framework, even if the mean structure can be partially fitted, small model biases accumulate over
411 long-term sequential recursion and are amplified in the terminal distribution. By reducing
412 reliance on global models through local conditional transition estimation, TGC is more robust
413 under such dynamic feedback structures^{3,28}. This result suggests that the primary applied value of
414 TGC may not lie in “arbitrary nonlinear” scenarios in general, but rather in settings that more
415 closely reflect the realities of longitudinal causal inference—namely, those with dynamic
416 feedback, observed confounding, and a focus on distributional shape in the terminal outcome
417 ^{28,29}.

418 Fourth, in the nonlinear heteroscedastic scenario, both methods exhibited substantial errors, and
419 TGC did not achieve comprehensive superiority over the parametric baseline. This result is
420 particularly important because it reminds us that nonparametric does not automatically imply
421 greater accuracy. The current TGC estimator relies on local nearest neighbors, kernel
422 smoothing, and OT couplings; while more flexible than global parametric models, it is also more
423 susceptible to the influence of neighborhood selection, sample sparsity, and high-dimensional
424 distance metrics^{6,19}. In strongly nonlinear and markedly heteroscedastic scenarios, local
425 neighborhoods may be unable to simultaneously balance bias and variance, leading to estimates
426 that still deviate from the true WCE^{19,23}. That is, the methodological advantage of TGC lies more
427 in “reducing the risk of model misspecification” rather than guaranteeing optimal numerical
428 performance in all complex scenarios⁶.

429 From the perspective of computational cost, the overhead of TGC is substantially higher than
430 that of parametric g-computation, which also warrants serious discussion. At each time point and
431 for each treatment arm, TGC must perform local neighborhood construction and OT solving,
432 resulting in a running time approximately two orders of magnitude longer than that of the
433 parametric method. This means that TGC is more suitable as a tool for small-to-medium-scale
434 data analysis or methodological research, rather than as a default option for large-scale high-
435 throughput analysis at its current stage. To further enhance its practicality, future work may
436 consider more efficient approximate OT algorithms, stratified sampling, sparse coupling, or
437 stronger parallelization implementations^{30,31}.

438 From a statistical perspective, our results also illustrate a more general issue: distributional
439 causal effects are not necessarily easier to estimate than mean-level causal effects. Although the
440 WCE provides a richer characterization of post-intervention changes, it simultaneously requires
441 estimation of the entire terminal distribution, and is therefore more sensitive to errors in each
442 step of the conditional transition^{9,10}. Especially under a longitudinal recursive structure, early-
443 stage errors propagate through successive layers, rendering the method sensitive to local
444 smoothing strategies, neighborhood size, sampling noise, and positivity conditions³.
445 Consequently, the advantage of the WCE lies primarily in its closer alignment with scientific
446 questions concerning “overall distributional change”⁹, rather than any inherent ease of
447 estimation⁶.

448 Our results also suggest several clear directions for methodological improvement. First, TGC
449 requires a more systematic hyperparameter selection strategy, particularly joint tuning of the
450 nearest-neighbor number, kernel bandwidth, and Sinkhorn regularization term¹⁶. Second, one
451 could consider automatic degeneration to a parametric branch in simple linear scenarios, or

adaptive switching between parametric and nonparametric estimation through model selection mechanisms. Third, for high-dimensional state variables, the current Euclidean distance may be insufficient to characterize true neighborhood structure; future work could explore learned distance metrics, representation learning, or OT estimation after dimensionality reduction¹⁴. Fourth, positivity deficiency and sample sparsity remain central challenges in dynamic causal inference; although TGC avoids spurious extrapolation through an explicit failure mechanism, this also implies that its applicability depends strongly on the support of the data. How to balance conservativeness with coverage is a question worthy of further investigation²⁷.

Overall, the simulation results in this paper yield a clear yet cautiously optimistic conclusion: TGC is not a method that outperforms traditional parametric g-computation in all scenarios, but it demonstrates better robustness in complex longitudinal confounding structures with dynamic feedback, and provides a natural approach to distributional causal effect estimation. Compared with traditional mean-based longitudinal causal analysis, TGC is better suited to settings where the post-intervention outcome changes not only in mean but also in distributional shape in a substantively meaningful way. If its computational efficiency and hyperparameter robustness can be further improved in the future, TGC is expected to become a competitive tool for distributional longitudinal causal inference.

Conclusions

Transport G-Computation provides a distributional framework for longitudinal causal inference by combining local conditional transition estimation with optimal transport. The simulation results suggest that the method is most relevant when treatment assignment and longitudinal state

473 evolution exhibit complex feedback, while computational cost, hyperparameter selection, and
474 local support diagnostics remain important considerations for practical use.

475 **List of abbreviations**

476 TGC: Transport G-Computation

477 WCE: Wasserstein causal effect

478 OT: Optimal transport

479 W2: Second-order Wasserstein distance

480 RMSE: Root mean square error

481 MAD: Mean absolute deviation

482 KL: Kullback-Leibler divergence

483 NNKCDE: Nearest-neighbor kernel conditional density estimation

484 **Declarations**

485 **Ethics approval and consent to participate**

486 Not applicable. This study uses simulated data only.

487 **Consent for publication**

488 Not applicable.

489 **Availability of data and materials**

490 The simulation code and generated data are available from the corresponding author upon
491 reasonable request.

492 **Competing interests**

493 The authors declare that they have no competing interests.

494 **Funding**

495 No specific funding was reported for this study.

496 **Authors' contributions**

497 YYH and XJH conceived the study and developed the overall research idea. YYH drafted the
498 manuscript and contributed to the design of the simulation study. TPF contributed to the
499 methodological development, computational implementation, and simulation analysis. JZ
500 contributed to the study design, statistical interpretation, and critical revision of the manuscript.
501 XXS contributed to the optimal transport methodology, interpretation of simulation results, and
502 manuscript revision. XJH supervised the study, provided clinical and methodological guidance,
503 and revised the manuscript critically for important intellectual content. All authors read and
504 approved the final manuscript.

505 **Acknowledgements**

506 Not applicable.

References

- [1] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period. *Mathematical Modelling*. 1986;7(9-12):1393-1512.
- [2] Hernan MA, Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
- [3] Gill RD, Robins JM. Causal inference for complex longitudinal data. *The Annals of Statistics*. 2001;29(6):1785-1811.
- [4] Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. *Statistics in Medicine*. 2013;32(9):1584-1618.
- [5] Vansteelandt S, Keiding N. G-computation—lost in translation? *American Journal of Epidemiology*. 2011;173(7):739-742.
- [6] Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology*. 2021;190(12):2730-2736.
- [7] Chatton A, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*. 2020;10(1):8549.
- [8] Chernozhukov V, Fernandez-Val I, Melly B. Inference on counterfactual distributions. *Econometrica*. 2013;81(6):2205-2268.
- [9] Bitler MP, Gelbach JB, Hoynes HW. What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review*. 2006;96(4):988-1032.

- 528 [10] Panaretos VM, Zemel Y. Statistical aspects of Wasserstein distances. *Annual Review of*
529 *Statistics and Its Application*. 2019;6:405-431.
- 530 [11] Chewi S, Niles-Weed J, Rigollet P. Statistical Optimal Transport. *Lecture Notes in*
531 *Mathematics* 2364. Springer; 2025.
- 532 [12] Naimi AI, Cole SR, Kennedy EH. An introduction to G methods. *International Journal of*
533 *Epidemiology*. 2017;46(2):756-762.
- 534 [13] Del Moral P, Doucet A, Singh SS. Uniform stability of a particle approximation of the
535 optimal filter derivative. *SIAM Journal on Control and Optimization*. 2015;53(3):1278-1304.
- 536 [14] Yang H, Tabak EG. Conditional density estimation, latent variable discovery, and optimal
537 transport. *Communications on Pure and Applied Mathematics*. 2022;75(3):610-663.
- 538 [15] Tabak EG, Trigila G, Zhao W. Conditional density estimation and simulation through
539 optimal transport. *Machine Learning*. 2020;109(4):665-688.
- 540 [16] Baptista R, Hosseini B, Kovachki N, et al. Conditional simulation via entropic optimal
541 transport: toward non-parametric estimation of conditional Brenier maps. arXiv:2411.07154.
542 2024.
- 543 [17] Villani C. *Optimal Transport: Old and New*. Springer; 2008.
- 544 [18] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of
545 average treatment effects. *Biometrika*. 2009;96(1):187-199.
- 546 [19] Izbicki R, Lee AB, Pospisil T. NNKCDE: Nearest Neighbor Kernel Conditional Density
547 Estimation. *Astronomy and Computing*. 2020;30:100362.

548 [20] Westreich D, Cole SR. Invited commentary: positivity in practice. *American Journal of*
549 *Epidemiology*. 2010;171(6):674-677.

550 [21] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in*
551 *Neural Information Processing Systems (NeurIPS)*. 2013;26:2292-2300.

552 [22] Stone CJ. Optimal global rates of convergence for nonparametric regression. *Annals of*
553 *Statistics*. 1982;10(4):1040-1053.

554 [23] Cheridito P, Eckstein S. Optimal transport and Wasserstein distances for causal models.
555 *Bernoulli*. 2025;31(2):1105-1139.

556 [24] Fan J, Yao Q, Tong H. Estimation of conditional densities and sensitivity measures in
557 nonlinear dynamical systems. *Biometrika*. 1996;83(1):189-206.

558 [25] Tabak EG, Trigila G, Zhao W. Data driven conditional optimal transport. *Machine*
559 *Learning*. 2021;110(11):3135-3155.

560 [26] Meyn SP, Tweedie RL. *Markov Chains and Stochastic Stability*. 2nd ed. Cambridge
561 University Press; 2009.

562 [27] Bao H, Schomaker M. Addressing Positivity Violations in Continuous Interventions through
563 Data-Adaptive Strategies. arXiv:2502.14566. 2025.

564 [28] Helske J. Feedback-augmented Non-homogeneous Hidden Markov Models for Longitudinal
565 Causal Inference. arXiv:2503.16014. 2025.

566 [29] Bica I, Alaa AM, Jordon J, van der Schaar M. Estimating counterfactual treatment outcomes
567 over time through adversarially balanced representations. *International Conference on Learning*
568 *Representations (ICLR)*. 2020.

- 569 [30] Altschuler J, Weed J, Rigollet P. Near-linear time approximation algorithms for optimal
570 transport via Sinkhorn iteration. *Advances in Neural Information Processing Systems (NeurIPS)*.
571 2017;30:1964-1974.
- 572 [31] Genevay A, Cuturi M, Peyre G, Bach F. Stochastic optimization for large-scale optimal
573 transport. *Advances in Neural Information Processing Systems (NeurIPS)*. 2016;29.