Robust Causal Inference for Large-Scale Healthcare Cost Control and Clinical Efficiency Optimization

Sining Chai,Tianyu Yang
Northeastern University, Boston, MA, USA

Abstract:To address the core contradiction of "rising costs and insufficient efficiency" in the healthcare system, this study proposes a healthcare system optimization framework integrating robust causal inference and large-scale optimization, targeting issues such as high-dimensional interference, hidden confounding, and lack of model robustness in observational healthcare data. First, a Robust Causal Inference Model (RCI-Model) is constructed, which eliminates hidden confounding through spectral transformation debiasing and identifies core associations via multi-scale causal graph pruning to achieve accurate estimation of clinical causal effects. Furthermore, with causal effects as constraints, an integer programming model incorporating efficacy and resource limitations is established, and cross-institutional resource optimization is completed by combining Lagrangian duality and federated learning. Experiments based on 530,000 hospitalization samples show that the average hospitalization cost is reduced by 18.7%, the bed turnover rate is increased by 23.4%, and the total regional healthcare cost across institutions is reduced by 15.2%. The research confirms that robust causal inference can separate spurious correlations in healthcare data, providing a reliable basis for cost control and efficiency optimization, and technical support for the implementation of hierarchical diagnosis and treatment.

Keywords: Robust Causal Inference; Large-Scale Optimization; Healthcare Cost Control; Clinical Efficiency; Hidden Confounding Elimination

1. Introduction

Population aging, upgraded health demands, and medical technology iteration have driven the global rise in healthcare expenditures, with the proportion of China's total health expenditure in GDP having increased significantly. However, the uneven allocation of medical resources and complex clinical processes lead to low efficiency. The average length of stay in top-tier hospitals is much longer than that in developed countries, and the imbalance between cost and efficiency has aggravated the burden on residents, becoming a core bottleneck in the implementation of the "Healthy China 2030" strategy. Traditional correlation analysis struggles to distinguish between causal and accompanying relationships, resulting in policies that "treat the symptoms but not the root cause." In contrast, robust causal inference can resist data noise and model bias, accurately identifying cost drivers and efficiency pathways. Currently, the application of robust causal inference in the healthcare field is still in its infancy. Although there have been explorations in policy evaluation and chronic disease treatment, it has not touched on the field of cost and efficiency. Additionally, there are common issues such as lack of robustness in research methods, separation of cost and efficiency studies, and limited data scale. Based on this, the multi-method integration innovation of this study has important practical value.

2. Theoretical Foundation and Method Construction

2.1 Core Theories of Robust Causal Inference

The core of causal inference is to identify the potential causal effect between intervention variables (e.g., treatment plans) and outcome variables (e.g., cost, efficacy), i.e., calculating the Average Treatment Effect (*ATE*): $ATE=E[Y(1)-Y(0)]$, where $Y(1)$ and $Y(0)$ represent the outcomes under the intervention and control states, respectively. Hidden confounding U in healthcare data will cause bias in traditional estimators, i.e., $E[Y|T=1]-E[Y|T=0]=ATE+E[U|T=1]-E[U|T=0]$,

where T is the intervention variable.

Robustness design needs to meet dual objectives: first, eliminating estimation bias caused by hidden confounding, and second, reducing the model's sensitivity to data distribution shifts. This study adopts a two-stage strategy of "spectral transformation debiasing + causal graph pruning": in the first stage, the leading singular values of the design matrix X are compressed through a spectral transformation matrix Q, reducing the perturbation term $//X_b//^2$ to a negligible level (optimal effect when ρ=0.5); in the second stage, confounding paths are identified based on Directed Acyclic Graphs (DAG), and spurious association edges are deleted using the IG pruning method to retain the direct causal path of "intervention-mediator-outcome" [1,2].

2.2 Large-Scale Optimization Model Based on Causal Effects

With the causal effects identified by the RCI-Model as constraints, an integer programming model for healthcare resource optimization is constructed. The optimization objective is to minimize the total regional healthcare cost while satisfying efficacy constraints and resource capacity constraints:

Objective function: $\min Z=\sum\sum(c_i x_i+d_i y_i)$

where c_i is the direct cost of the i-th type of patient receiving the j-th treatment plan, d_i is the indirect cost (e.g., bed occupancy fee), x_i is the number of patients allocated, and y_i is the input of supporting resources.

Constraints: 1) Causal efficacy constraint: Based on the treatment effect estimated by the RCI-Model, the cure rate of the i-th type of patient after receiving plan j must be $\geq\theta_i$ (θ_i is set according to the disease type, e.g., $\theta_i\geq$0.6 for cancer patients); 2) Resource capacity constraint: $\sum x_i\leq C$ (C is the equipment capacity of the j-th plan); 3) Hierarchical diagnosis and treatment constraint: The proportion of consultations in primary medical institutions $\geq\alpha$ (α=0.45 , in line with national policy requirements).

The Lagrangian duality method is used to convert the integer programming into a convex optimization problem, which is solved by the Alternating Direction Method of Multipliers (ADMM). The convergence time is controlled within 2.5 hours for 100,000-level samples, meeting the timeliness requirements of clinical decision-making.

2.3 Cross-Institutional Collaborative Optimization Mechanism

To address the problem of healthcare data "silos," a federated learning framework is introduced: each medical institution acts as a client, training local parameters of the RCI-Model locally and only uploading model update gradients to the cloud server; the server generates a global model by aggregating gradients (using weighted average, where weights are proportional to the institution's sample size) and then distributes it to each client for parameter fine-tuning. Homomorphic encryption technology is used to encrypt gradients to ensure patient privacy and security, with the model performance loss controlled within 5%.

3. Experimental Verification and Result Analysis

3.1 Experimental Data and Scenario Design

Experimental data are sourced from three channels: 1) A provincial DRG database (2022-2024), containing 530,000 hospitalization samples, covering clinical records, cost details, and prognostic information of 28 diseases; 2) Lung cancer gene expression data from the TCGA database, including methylation data of 865,860 CpG sites from 1,200 patients; 3) 24 independent validation sets from the GEO database for testing model generalization ability [3,4,5].

Three experimental scenarios are set: 1) Single-center cost optimization (targeting the Cardiology Department of a Grade A tertiary hospital); 2) Cross-institutional collaborative optimization (covering a hierarchical diagnosis and treatment network of 12 hospitals); 3) Robustness test (simulating data distribution shifts, such as adding low-quality data from primary hospitals). Comparative methods include traditional regression analysis (OLS), Propensity Score Matching (PSM), Debiased Lasso (DL), and DDL methods.

3.2 Evaluation Index System

A three-dimensional evaluation system of "causal accuracy - optimization effect - robustness" is constructed:

Causal accuracy: Estimation bias (BIAS), Root Mean Square Error (RMSE), and E-value (measuring confounding resistance, E>1.5 is

excellent);
Optimization effect: Average hospitalization cost, bed turnover rate, and treatment effective rate;
Robustness: Model performance degradation rate after data distribution shift (ΔACC), and cross-institutional model adaptation error.

3.3 Experimental Results and Analysis

3.3.1 Comparison of Causal Inference Performance

In the lung cancer clinical dataset, the causal effect estimation results of various methods are shown in Table 1. The estimation bias of the RCI-Model (-0.023) is only 11.6% of that of the PSM method (-0.198), the RMSE is 18.7% lower than that of the DDL method, and the E-value reaches 1.92, significantly higher than other methods, indicating its advantage in eliminating hidden confounding. This result is consistent with the research conclusions of Fudan University in the ADNI database, verifying the effectiveness of spectral transformation debiasing technology in biomedical data [6].

Table 1. Comparison of Causal Inference Performance of Various Methods

Method	Estimation Bias (BIAS)	Root Mean Square Error (RMSE)	E-value
OLS	-0.245	0.312	1.03
PSM	-0.198	0.256	1.21
DL	-0.087	0.093	1.56
DDL	-0.052	0.092	1.68
RCI-Model	-0.023	0.075	1.92

3.3.2 Cost Control and Efficiency Optimization Effects

In the single-center scenario, the optimization scheme based on the RCI-Model reduces the average hospitalization cost of the Cardiology Department from 18,620 yuan to 15,180 yuan, a decrease of 18.7%; the bed turnover rate increases from 1.2 times/month to 1.48 times/month, an increase of 23.4%; the treatment effective rate remains at 92.3%, an increase of 1.2 percentage points compared with before optimization. The cost reduction mainly comes from two aspects: first, eliminating "high-cost and low-efficacy" treatment plans through causal identification (e.g., the usage rate of a certain imported antibiotic decreases from 32% to 11%); second, optimizing the inspection process, reducing 37% of duplicate inspection items.

In the cross-institutional scenario, after 12 hospitals constructed a collaborative optimization model through federated learning, the total regional healthcare cost decreased by 15.2%, and the proportion of consultations in primary medical institutions increased from 38% to 46%, achieving the hierarchical diagnosis and treatment goal of "severe diseases treated in hospitals, minor illnesses treated in communities." This result verifies the practicality of the causal optimization model in resource allocation, and its effect is superior to resource allocation methods based solely on administrative orders.

3.3.3 Model Robustness Test

In the data distribution shift test, after adding 20% low-quality data from primary hospitals (containing more noise and missing values) to the training set, the performance changes of various methods are shown in Table 2. The ΔACC of the RCI-Model is only 8.7%, significantly lower than that of the DL method (23.5%) and the PSM method (31.2%), indicating that through the dual mechanism of spectral transformation and causal graph pruning, it effectively reduces the interference of noisy data on the model. In terms of cross-institutional adaptation error, the prediction error of the RCI-Model in new hospitals is only 0.062, meeting the requirements of clinical decision-making [7,8,9].

Table 2. Comparison of Model Robustness After Data Distribution Shift

Method	Performance Degradation Rate (ΔACC)	Cross-Institutional Adaptation Error
OLS	42.3%	0.187
PSM	31.2%	0.145
DL	23.5%	0.098
DDL	15.6%	0.081
RCI-Model	8.7%	0.062

4. Strategies and Recommendations

4.1 Cost Control Strategies Based on Robust Causal Inference

In terms of cost control, it is necessary to break through the limitations of traditional administrative control and construct a four-dimensional precision strategy. First, precise price regulation: relying on the robust causal model to define a reasonable cost range, establishing a national healthcare cost database, and using robust quantile regression to

distinguish cost differences between eastern, central, and western regions and hospitals of different levels; adopting a "cost-plus + efficacy evaluation" model for innovative drugs, balancing R&D incentives and patient burden through the robust difference-in-differences model, and controlling the circulation markup rate within 15% through robust outlier detection. Second, differentiated DRG/DIP policies: classifying diseases according to the results of robust cluster analysis, increasing the payment standard for complex diseases by 15%-20% with a severe disease supplementary mechanism, providing a 10% payment premium for primary hospitals accepting common diseases, and dynamically evaluating and adjusting every quarter using robust DID. Third, full-process cost supervision: establishing a "cost-efficacy" correlation model, identifying ineffective costs using robust propensity score matching; promoting the integrated model of clinical pathways and cost control, with supporting robust causal-driven intelligent medical insurance review. Fourth, regional collaborative cost control: identifying core factors of regional costs through the robust spatial econometric model, monitoring over-treatment in eastern regions and optimizing procurement costs in western regions, forming cross-regional procurement alliances to reduce the price of medical consumables, while prioritizing talent training in western regions and technology transfer from eastern regions to improve the equipment utilization rate of county-level hospitals in western regions.

4.2 Recommendations for Clinical Efficiency Optimization Based on Robust Causal Inference

Optimizing clinical efficiency needs to focus on core bottlenecks and implement a four-dimensional improvement plan. In process restructuring, priority should be given to breaking through the bottlenecks of inspection appointment and result transmission, building a regional centralized appointment platform to reduce the inspection appointment time from 2.3 days to 0.5 days, and realizing cross-hospital mutual recognition of inspection results through blockchain technology; implementing "general practitioner-specialist" joint clinics to shorten referral time. In resource sinking, implement a three-dimensional strategy of "talent-technology-patients": train 5,000 general practitioners annually in western regions, ensuring that the salary of primary care doctors is not lower than the average level of local public institutions; require doctors in Grade A tertiary hospitals to conduct at least 4 days of outpatient services in primary institutions every month, coupled with differences in medical insurance reimbursement incentives to increase the primary diagnosis rate accordingly. Technology empowerment should be promoted in phases according to the priority of causal effects: achieve full coverage of electronic medical record interoperability and teleconsultation in county-level hospitals within 1-2 years, promote AI-assisted diagnosis technology within 3-4 years, and construct an intelligent clinical platform in the long term, with effects evaluated using robust models every six months. In the incentive mechanism, take "average daily effective clinical time (30%), patient satisfaction (25%), readmission rate (25%), and cost control rate (20%)" as core indicators, focusing on satisfaction in primary institutions and the volume of complex case diagnosis and treatment in Grade A tertiary hospitals [10,11].

4.3 Implementation Guarantee Measures

To ensure the implementation of the strategies, it is necessary to consolidate the four major guarantees of policy, technology, talent, and ethics. At the policy level, issue the "Specifications for Healthcare Data Quality Management," establish a robust evaluation system for medical policies, and incorporate causal analysis into hospital grade evaluation. In technological innovation, build a national-level healthcare big data center and adopt federated learning to ensure security. In talent training, add healthcare data analysis majors in universities, train 50,000 in-service personnel annually, and introduce high-end experts. Ethically, establish a data use review mechanism, implement de-identification processing and patient informed consent systems, and balance data utilization and privacy protection. These measures will promote the in-depth application of robust causal inference and assist the high-quality development of the healthcare industry.

5. Conclusion and Outlook

The Robust Causal Inference Model (RCI-Model) proposed in this study effectively solves

the hidden confounding problem in high-dimensional healthcare data through the dual mechanism of spectral transformation debiasing and multi-scale causal graph modeling, and its causal effect estimation accuracy and robustness are superior to existing methods. The large-scale optimization framework based on this model reduces the average hospitalization cost by 18.7%, increases the bed turnover rate by 23.4% in 530,000 medical samples, and shows good adaptability in cross-institutional scenarios, verifying the feasibility and superiority of the integration of causal inference and optimization algorithms in healthcare systems. The research indicates that robust causal inference can provide reliable technical support for healthcare cost control and clinical efficiency optimization, and its integration with federated learning provides a new solution for the implementation of hierarchical diagnosis and treatment policies, with important theoretical value and clinical application prospects.

Future research can focus on the following three aspects: first, improving model generalization: developing nonlinear spectral transformation technology based on deep learning to adapt to complex nonlinear healthcare scenarios such as tumor progression and chronic disease management, breaking through the application limitations of current linear models; second, constructing dynamic optimization capabilities: integrating reinforcement learning algorithms to capture dynamic factors such as public health emergencies and medical policy adjustments in real time, achieving agile response in resource allocation; third, improving the clinical implementation system: collaborating with medical institutions to build a closed-loop mechanism of "model training - clinical pilot - effect feedback," optimizing model parameters for special departments such as pediatrics and emergency, and promoting the standardization and localization of robust causal analysis toolkits to reduce the application threshold for primary medical institutions, helping modern healthcare benefit a wider group.

References

[1]Ghosh A, Rothenhäusler D. Assumption-robust Causal Inference[J]. arXiv preprint arXiv:2505.08729, 2025.

[2]Cho E, Yang S. Variable selection for doubly robust causal inference[J]. Statistics and its interface, 2024, 18(1): 93.

[3]Xiao T, Wang S. Towards unbiased and robust causal ranking for recommender systems[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1158-1167.

[4]Wang J W, Dai M W, Liang P, et al. Correlation does not equal causation: the imperative of causal inference in machine learning models for immunotherapy[J]. Frontiers in Immunology, 2025, 16: 1630781.

[5]Zhang T, Shan H R, Little M A. Causal GraphSAGE: A robust graph method for classification based on causal sampling[J]. Pattern recognition, 2022, 128: 108696.

[6]Sanchez P, Voisey J P, Xia T, et al. Causal machine learning for healthcare and precision medicine[J]. Royal Society Open Science, 2022, 9(8): 220638.

[7]Zhang Y, Chen S, Liu D. A measurement study of the environmental quality and medical expenditures of elderly individuals: causal inference based on machine learning. Arch Public Health. 2024 Oct 30;82(1):195.

[8]Raita Y, Camargo Jr C A, Liang L, et al. Big data, data science, and causal inference: a primer for clinicians[J]. Frontiers in Medicine, 2021, 8: 678047.

[9]Hu L, Gu C. Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes[J]. Health Services and Outcomes Research Methodology, 2021, 21(3): 287-308.

[10]Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects[J]. Clinical epidemiology, 2018: 771-788.

[11]Irankhah E, Pagare M, Chetla L, et al. Machine learning-enhanced causal inference of surgical decisions and rehabilitation strategies in traumatic brain injury[J]. Frontiers in Neurology, 2025, 16: 1685335.

