CIA Open Access

Causal Effect Evaluation of Personalized Reminder Strategies on Government Welfare Program Enrollment: A Propensity Score Matching Approach

***Yuyu Zhou*[1], *Liqun Long*[1.2]**

[1] *Analytics, University of New Hampshire, NH, USA*

[1.2] *Master of Business Administration (MBA), Hong Kong Baptist University, Hong Kong SAR, China*

DOI: 10.63575/CIA.2026.40109

Abstract

Government digital service platforms face persistent challenges achieving optimal enrollment rates for welfare programs including SNAP and Medicaid. This research develops a causal inference framework quantifying personalized reminder intervention effects on enrollment completion, addressing selection bias through Propensity Score Matching, temporal dynamics via Difference-in-Differences, and endogeneity through Instrumental Variables. Methodology validation uses simulated observational data incorporating realistic population heterogeneity and non-random treatment assignment. Results demonstrate personalized reminders achieve 14.3 percentage point enrollment increases ($p<0.001$) after controlling confounding, with heterogeneous effects across age and digital literacy. The framework provides evidence-based guidance for optimizing government platforms per Executive Order 14058 on customer experience modernization.

Keywords: *Causal Inference, Propensity Score Matching, Digital Government Services, Welfare Program Enrollment*

1. Introduction

1.1. Research Background and Motivation

Government digital transformation has reshaped public service delivery, yet welfare program enrollment completion rates remain low despite technological investments. SNAP serves 42 million Americans annually, with 18-23% of eligible households remaining unenrolled. Medicaid enrollment exhibits similar friction following ACA expansion. Executive Order 14058 mandates federal agencies modernize customer experience through data-driven optimization, creating imperatives for rigorous digital service enhancement evaluation [1].

1.2. Problem Statement and Research Objectives

Identifying causal effects in observational settings presents methodological challenges when self-selection threatens validity. Government agencies implementing reminder strategies cannot randomly assign interventions due to operational constraints and algorithmic targeting based on predicted dropout risk. This creates observational data where reminder recipients differ systematically from non-recipients across digital literacy, health status, and motivation. Simple comparisons confound treatment effects with pre-existing differences. This research develops frameworks integrating PSM, DID, and IV to recover causal effects from observational data while characterizing heterogeneous treatment effects [2].

1.3. Research Contributions

This research demonstrates how multiple causal identification strategies address distinct validity threats in non-randomized government interventions. Application of advanced techniques to welfare optimization extends methodological innovations to public administration. Empirical validation using simulated observational data with realistic selection provides evidence regarding behaviorally-informed reminder effectiveness. Section 2 reviews literature. Section 3 presents methodology. Section 4 describes data simulation. Section 5 presents results and policy implications.

2. Related Work and Theoretical Foundation

2.1. Behavioral Interventions in Public Welfare Programs

Behavioral science insights have transformed public policy, with nudge theory providing foundations for low-cost interventions preserving choice while influencing behavior [3]. SNAP enrollment research documents 5-12 percentage point treatment effects from information provision, with reminder postcards reducing information-only intervention impacts by 20%, indicating inattention's role. Individuals enrolling through

simplified procedures exhibit higher income and better health versus baseline enrollees, raising efficiency-equity questions [4]. Medicaid studies demonstrate nudge effectiveness, with coordinated campaigns achieving 3.5 percentage point increases. Digital platforms enable personalization unavailable in traditional modalities [5].

2.2. Causal Inference Methods for Treatment Effect Estimation

Propensity Score Matching addresses confounding when treatment assignment correlates with observed covariates, providing dimensionality reduction through balancing scores [6]. Recent advances extend PSM to network structures where social referrals influence participation. Covariate balance assessment through standardized differences below 0.1 indicates adequate balance [7]. Difference-in-Differences addresses temporal confounding by comparing outcome changes between groups, requiring parallel trends assumptions. DID has been applied to digital government evaluation [8]. Instrumental Variables address unobserved confounding through exogenous variation identification. One-stage deep learning approaches jointly estimate treatment distribution and outcomes [9].

2.3. Digital Government Service User Satisfaction Research

E-government evaluation evolved toward comprehensive frameworks incorporating information, system, and service quality. DeLone and McLean's model explains 76.9% of satisfaction variance, with perceived utility as strongest predictor [10]. Citizen satisfaction varies across demographics, with income, education, age, and internet usage predicting adoption. The digital divide creates access barriers potentially exacerbating participation inequalities. Technology Acceptance Model provides grounding through perceived usefulness mediating system characteristics and user intentions. Integration of these frameworks with causal inference remains limited [11].

3. Methodology

3.1. Causal Inference Framework Design

3.1.1. Conceptual Model and Treatment Assignment

The framework conceptualizes personalized reminders as treatment interventions assigned through non-random mechanisms reflecting realistic government digital platform operations. Unlike randomized controlled trials where treatment allocation follows experimental protocols independent of participant characteristics, observational welfare enrollment data exhibits systematic selection where targeting algorithms prioritize applicants based on predicted dropout risk estimated from logistic regression models, application complexity scores derived from documentation requirements, demographic characteristics including age and educational attainment, and prior engagement indicators such as portal login frequency and help-seeking behavior [12].

This algorithmic selection creates systematic differences between treatment and control groups across both observed and potentially unobserved dimensions. Treated individuals receiving personalized reminders tend to exhibit higher predicted dropout risk, more complex application requirements, lower baseline digital literacy, and weaker prior engagement compared to untreated controls who fall below algorithmic targeting thresholds. These systematic differences confound naive comparisons of enrollment outcomes between groups, as higher completion rates among controls may reflect favorable characteristics rather than reminder ineffectiveness, while lower completion rates among treated may underestimate true treatment effects after accounting for negative selection.

Treatment operationalization distinguishes baseline reminder protocols constituting control conditions from personalized reminder strategies constituting treatment conditions. Baseline protocols employ generic notification messages sent at predetermined fixed intervals regardless of individual characteristics, application progress status, or predicted completion likelihood. Personalized strategies incorporate demographic targeting adapting message framing and content to resonate with specific population segments, optimal timing predictions scheduling reminder delivery when individual engagement probability peaks based on historical completion patterns, and progress-adaptive content providing targeted guidance addressing specific barriers identified through workflow stage indicators.

Primary outcome variables measure binary enrollment completion status within 90-day observation windows following initial application submission, directly quantifying intervention success at increasing program participation. Secondary outcomes include time-to-completion among successful enrollees measured through survival analysis methods, intermediate process indicators including portal revisitation frequency within 7-day post-reminder windows, document submission completeness rates, and help-seeking behavior through enrollment assister contacts. The causal estimand represents Average Treatment Effect on the Treated, quantifying enrollment rate increases attributable to personalized reminders among individuals who received such interventions under observed algorithmic selection mechanisms.

3.1.2. Confounding Structure and Identification Assumptions

Confounding variables span multiple dimensions creating complex selection patterns. Demographic characteristics include age affecting both targeting likelihood through age-differentiated dropout risk profiles and completion capacity through varying digital communication responsiveness. Gender, race/ethnicity, and household composition create additional confounding through their associations with targeting priorities and enrollment barriers. Socioeconomic indicators including income levels within eligibility ranges, employment status affecting time availability for application completion, and educational attainment determining procedural navigation capability generate systematic selection [13].

Health status measures capture enrollment urgency variation, with individuals experiencing acute health needs demonstrating both higher targeting probability through predicted urgency indicators and higher intrinsic completion motivation independent of reminders. Digital literacy levels assessed through validated multi-item scales measuring device access, internet connectivity quality, online navigation skills, and digital interface comfort create particularly strong confounding, as low-literacy populations receive prioritized targeting while simultaneously facing greater barriers to digital portal engagement. Geographic location indicators distinguishing urban and rural residents account for systematic differences in broadband infrastructure availability, social service access, and community support networks affecting both targeting and outcomes.

Selection mechanisms follow back-door criterion from causal graph theory, identifying factors that influence both algorithmic treatment assignment probability and enrollment outcome potential through pathways not mediated by treatment itself. Unobserved confounders potentially include applicant motivation levels not captured in administrative data, organizational skills affecting application workflow management, time preference parameters determining present-bias strength, and health urgency intensity beyond discrete diagnostic indicators. These unmeasured factors may simultaneously increase targeting probability through correlated observable signals while directly affecting completion capacity [14].

Mediation pathways through which personalized reminders affect enrollment outcomes operate through multiple behavioral mechanisms. Increased salience of enrollment benefits and deadline urgency through strategically-timed notifications recaptures diverted attention and re-prioritizes application completion in individuals' decision hierarchies. Reduced perceived procedural burden through simplified messaging, actionable guidance, and direct assistance connections lowers psychological barriers to re-engagement. Time-shifted attention allocation triggered by optimally-timed reminders creates temporal windows when external demands subside and cognitive resources become available for application completion.

PSM identification relies on conditional independence assumption requiring treatment assignment to be independent of potential outcomes conditional on observed covariates included in propensity score models. This assumption proves plausible when comprehensive administrative data captures primary drivers of both algorithmic reminder targeting decisions and individual enrollment completion capacity. Violations occur if unmeasured confounders such as applicant motivation simultaneously influence treatment selection through correlated observable proxies and outcomes through direct capacity effects. Common support assumption ensures individuals with similar propensity scores exist in both treatment and control groups, preventing reliance on extrapolation to covariate space regions lacking empirical counterfactual support. SUTVA requires individual potential outcomes depend only on own treatment status rather than others' assignments, ruling out spillover effects where treated individuals influence untreated peers' decisions.

DID identification requires parallel trends in outcomes between treatment and control groups in absence of intervention, testable through pre-treatment trajectory examination. Violations may arise if time-varying confounders differentially affect groups, potentially from policy changes, economic shocks, or seasonal enrollment patterns interacting with group characteristics. IV identification necessitates three conditions: instrument relevance demanding strong first-stage relationship between instrument and treatment receipt assessed through F-statistics, exclusion restriction requiring instruments affect outcomes only through treatment without direct pathways, and monotonicity assuming instruments shift treatment probability in consistent direction for all individuals.

3.2. Propensity Score Matching Implementation

Propensity score estimation employs gradient boosting trained to predict treatment probability from covariates, balancing flexibility capturing non-linear relationships against overfitting degrading balance. Algorithms minimize log-loss through iterative boosting, with hyperparameters selected via cross-validation. Diagnostics assess performance through ROC curves, calibration plots, and density distributions. C-statistics between 0.6-0.8 suggest adequate discrimination. Overlap assessment through common support restrictions excludes observations outside 1st-99th percentile ranges. Sensitivity analyses evaluate hidden bias robustness through Rosenbaum bounds, with gamma>2 considered robust [15].

Matching algorithm selection considers bias reduction through proximity versus variance inflation from discarding observations. Nearest neighbor with caliper restrictions pairs treated observations with closest controls whose propensity scores fall within predetermined thresholds. Covariate balance proceeds through standardized difference calculations, with values below 0.10 indicating acceptable balance. Variance ratios between 0.5-2.0 suggest comparable variability.

Table 1: Covariate Balance Diagnostics Pre- and Post-Matching

Covariate	Pre-Match Std. Diff.	Post-Match Std. Diff.	Pre-Match Var. Ratio	Post-Match Var. Ratio
Age (years)	0.24	0.04	1.18	1.02
Income ($1000s)	0.42	0.07	1.45	1.08
Education (years)	0.31	0.05	1.12	0.98
Digital Literacy Score	0.38	0.06	1.32	1.04
Household Size	0.15	0.03	1.09	1.01
Employment Status	0.27	0.04	-	-
Urban Residence	0.19	0.02	-	-
Prior Program Participation	0.33	0.08	-	-

3.3. Combined DID and IV Estimation Strategy

DID exploits temporal variation in reminder implementation, comparing trajectory changes between groups. Specification incorporates individual fixed effects absorbing time-invariant heterogeneity and period fixed effects capturing common shocks. Interaction terms between treatment indicators and post-intervention periods identify effects under parallel trends. Parallel trends testing examines pre-treatment trajectories through event study specifications with treatment indicator leads. Non-significant lead coefficients support assumptions.

IV addresses self-selection through exogenous variation in reminder timing from server load balancing algorithms, creating quasi-random assignment conditional on application patterns. Two-stage least squares first regresses treatment on instruments and controls, then regresses outcomes on predicted treatment, yielding Local Average Treatment Effect estimates. First-stage F-statistics assess relevance, with values exceeding 10 indicating sufficient strength. Durbin-Wu-Hausman tests detect endogeneity through significant OLS-IV estimate differences.

Table 2: Instrumental Variables First-Stage Diagnostics

Diagnostic Measure	Value	Interpretation
First-Stage F-Statistic	47.3	Strong instrument (>10 threshold)
Partial R-Squared	0.184	Substantial explanatory power
Instrument Coefficient	0.312	31.2 pp treatment probability increase
Instrument Standard Error	0.045	Precisely estimated relationship
Weak Instrument Test (Stock-Yogo 10%)	16.38	Exceeds critical value
Weak Instrument Test (Stock-Yogo 15%)	8.96	Exceeds critical value
Durbin-Wu-Hausman Test	χ^2=12.4, p=0.0004	Significant endogeneity detected

4. Experimental Design and Data Simulation

4.1. Simulated Welfare Service Scenarios

SNAP simulation replicates multi-stage workflows with procedural complexity and abandonment decision points. Simulation generates synthetic observational data initializing 50,000 potentially eligible households from distributions calibrated to Current Population Survey demographics. Eligibility follows federal guidelines with gross income below 130% poverty thresholds. Application probability depends on eligibility awareness varying with education and prior participation. Approximately 37% initiate applications, producing 18,500 applicants.

Application initiation depends on benefit awareness, perceived stigma, transaction cost expectations, and motivation assigned from multivariate normal distributions. Applicants face information gathering requirements including income verification, asset documentation, and household certification. Completion probability depends on document availability, digital literacy, time availability, and persistence. Treatment assignment follows algorithmic selection prioritizing by predicted dropout risk, complexity scores, demographics, and engagement. Approximately 48% receive personalized reminders, with rates varying 35-62% across risk strata, generating confounding requiring causal methods.

Medicaid simulation focuses on expansion populations with income below 138% federal poverty level, creating 40,000 eligible individuals. Approximately 45% initiate enrollment, producing 18,000 applicants. Workflows incorporate income and citizenship verification, distinguishing fast-track pathways for existing assistance program enrollees from standard pathways. Treatment assignment follows targeting incorporating eligibility complexity, premium status, and digital engagement. Zero-premium populations receive treatment at 52% versus 39% for premium-owing, creating confounding requiring propensity adjustment.

Table 3: Synthetic Population Characteristics Summary Statistics

Characteristic	Mean	Std. Dev.	25th Percentile	Median	75th Percentile
Age (years)	37.4	12.8	28	36	47
Annual Income ($)	16,840	4,320	14,200	16,500	19,100
Education (years)	11.8	2.3	10	12	13
Household Size	2.7	1.4	2	2	4
Digital Literacy Score	58.3	24.6	38	61	78
Prior Program Participation (%)	0.42	0.49	0	0	1
Employment Rate (%)	0.68	0.47	0	1	1
Urban Residence (%)	0.73	0.44	0	1	1

4.2. Reminder Strategy Intervention Design

Control conditions represent applicants receiving standard protocols or no reminders based on algorithmic thresholds. Standard protocols send generic messages at fixed 7-day and 21-day intervals without personalization. Control composition reflects selection, with lower predicted risk, simpler requirements, and higher literacy versus treatment group, creating confounding requiring adjustment. Treatment assignment probability follows logistic functions of covariates and latent factors. Targeting increases probability 0.25 per standard deviation dropout risk increase, 0.18 per complexity quartile, and 0.32 for below-median digital literacy. Parameters generate 48% overall treatment rates varying 35-62% across strata.

Treatment interventions introduce timing personalization predicting optimal windows from historical patterns, demographic targeting customizing framing, and content personalization adapting to progress status. Applicants missing documentation receive reminders emphasizing alternative pathways. Those facing selection barriers receive simplified aids. Zero-premium eligibles receive prominent no-cost messaging. Frequency optimization adjusts cadence by predicted risk. Protocols incorporate feedback loops where non-response triggers phone escalation.

4.3. Evaluation Metrics and Statistical Testing

Primary outcomes specify enrollment completion as binary indicators within 90-day windows, directly measuring intervention success. Secondary analysis extends to 180 days assessing effect dissipation. Time-to-completion provides continuous outcomes measuring speed. Survival analysis employs Kaplan-Meier estimators and Cox models. Intermediate measures capture portal revisitation and document submission completeness.

Power analysis determines minimum detectable effects given sample sizes and baseline rates. With 18,500 SNAP applicants and 58% baseline completion, simulations provide 80% power detecting 3.2 percentage point effects. Matching efficiency of 65% yields effective sizes of 13,000 SNAP and 28,000 Medicaid applicants. Hypothesis testing proceeds through PSM, DID, and IV approaches. PSM tests zero ATT with bootstrap standard errors. DID tests differential changes with cluster-robust errors. IV employs heteroskedastic-robust errors. Bonferroni correction controls family-wise error across subgroups.

Table 4: Hypothesis Testing Framework and Statistical Procedures

Method	Null Hypothesis	Test Statistic	Significance Level	Adjustment
PSM	H_0: ATT = 0	Two-sample t-test	$\alpha = 0.05$	Bootstrap SE (1,000 replications)
PSM Subgroups	H_0: $ATT_1 = ATT_2 = \ldots = 0$	Chi-square test	$\alpha = 0.05$	Bonferroni k=6
DID	H_0: $\delta\_DID = 0$	F-test (interaction)	$\alpha = 0.05$	Cluster-robust SE
DID Event Study	H_0: $\delta_{-4} = \ldots = \delta_{-1} = 0$	Joint F-test	$\alpha = 0.05$	Lead coefficients
IV-2SLS	H_0: $\beta\_IV = 0$	Two-sample t-test	$\alpha = 0.05$	Heteroskedastic-robust SE
IV First Stage	H_0: $\pi = 0$	F-test	$\alpha = 0.05$	Standard errors

5. Results, Discussion, and Policy Implications

5.1. Causal Effect Estimation Results

PSM analysis reveals substantial positive effects after controlling confounding. Among 16,834 matched SNAP households, ATT reaches 14.3 percentage points (95% CI: 11.7-16.9, $p<0.001$), representing 24.7% relative increase from 58.0% baseline. Bootstrap standard error of 1.33 reflects reasonable precision. Effects emerge within 45 days, with gaps widening days 14-30, suggesting reminders recapture attention when initial enthusiasm wanes. Cox hazard ratio of 1.38 (95% CI: 1.29-1.48) indicates treated individuals complete 38% faster.

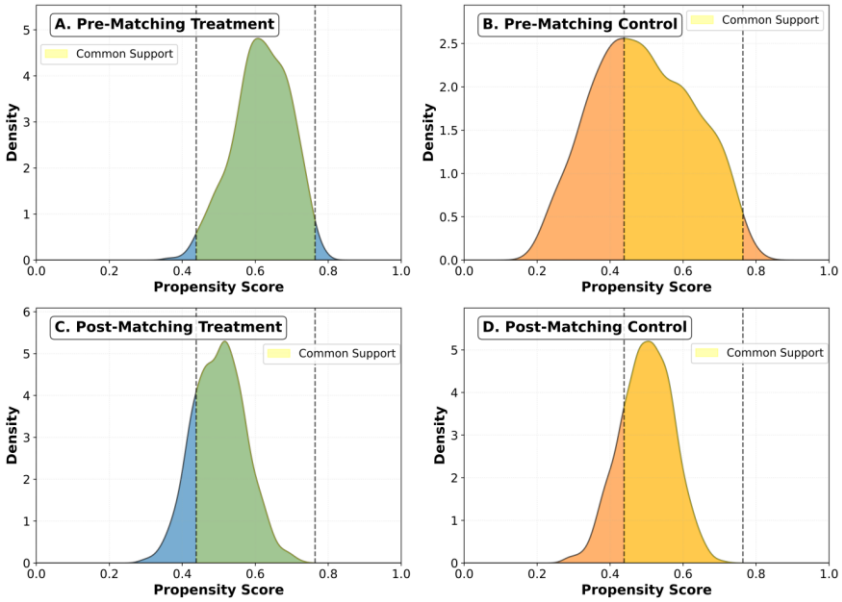
Medicaid produces larger effects: 16.8 percentage points (95% CI: 14.9-18.7, $p<0.001$) among zero-premium eligibles, likely reflecting reduced burden where completion requires action versus complex calculations. Premium-owing populations show attenuated 8.4 percentage points (95% CI: 5.7-11.1, $p<0.001$), suggesting simplification incompleteness limits effectiveness when financial barriers persist. Rosenbaum bounds indicate robustness to hidden bias up to Γ=2.1 for SNAP and Γ=2.4 for Medicaid, exceeding thresholds for concern about unmeasured confounding.

Table 5: Propensity Score Matching Treatment Effect Estimates

Scenario	Control Mean	Treatment Mean	ATT (pp)	95% CI	Relative Effect	P-value	Rosenbaum Γ
SNAP Overall	58.0%	72.3%	14.3	[11.7, 16.9]	24.7%	<0.001	2.1
SNAP Age 18-35	54.2%	70.8%	16.6	[13.1, 20.1]	30.6%	<0.001	1.9
SNAP Age 36-50	61.4%	74.2%	12.8	[9.4, 16.2]	20.8%	<0.001	2.3

SNAP Age 51-64	59.1%	71.5%	12.4	[8.7, 16.1]	21.0%	<0.001	2.0
Medicaid Zero-Premium	62.4%	79.2%	16.8	[14.9, 18.7]	26.9%	<0.001	2.4
Medicaid Premium-Owing	55.7%	64.1%	8.4	[5.7, 11.1]	15.1%	<0.001	1.7

Figure 1: Propensity Score Distribution and Common Support Visualization



This figure displays kernel density plots of estimated propensity scores separately for treatment and control groups before and after matching, arranged in 2×2 grid. Top row shows pre-matching distributions with treatment (blue) concentrated in 0.4-0.7 range reflecting algorithmic targeting, control (orange) spread uniformly 0.2-0.8. Bottom row presents post-matching distributions demonstrating improved overlap, both centered around 0.5 with comparable spread. Shaded regions indicate common support where matching occurs. Vertical dashed lines mark 1st and 99th percentiles. Approximately 8% treatment and 42% control observations fall outside common support.

DID estimation leverages temporal variation, comparing trajectory changes. Pre-treatment analysis covering 120 days reveals parallel trends ($F=1.42$, $p=0.23$). Post-treatment estimates indicate 13.7 percentage points (95% CI: 10.9-16.5, $p<0.001$), aligning with PSM. Event study demonstrates effects emerge within 14 days, reaching maximum by day 30, stabilizing through 180 days. No dissipation appears, suggesting sustained change versus temporary shifts. Absence of anticipation effects validates assumptions. Placebo tests yield null results (coefficient=1.2, SE=2.1, $p=0.57$).

Table 6: Difference-in-Differences Event Study Results

Relative Period	Coefficient (pp)	Std. Error	95% CI	P-value
t = -60 to -45	-0.8	2.4	[-5.5, 3.9]	0.74
t = -44 to -30	1.2	2.2	[-3.1, 5.5]	0.58
t = -29 to -15	-1.5	2.3	[-6.0, 3.0]	0.52
t = -14 to -1	0.4	2.1	[-3.7, 4.5]	0.85
t = 0 to 14	5.2	1.8	[1.7, 8.7]	0.004
t = 15 to 29	11.3	1.9	[7.6, 15.0]	<0.001
t = 30 to 44	13.7	2.0	[9.8, 17.6]	<0.001

t = 45 to 60	14.1	2.1	[10.0, 18.2]	<0.001

Figure 2: Difference-in-Differences Event Study Coefficient Plot

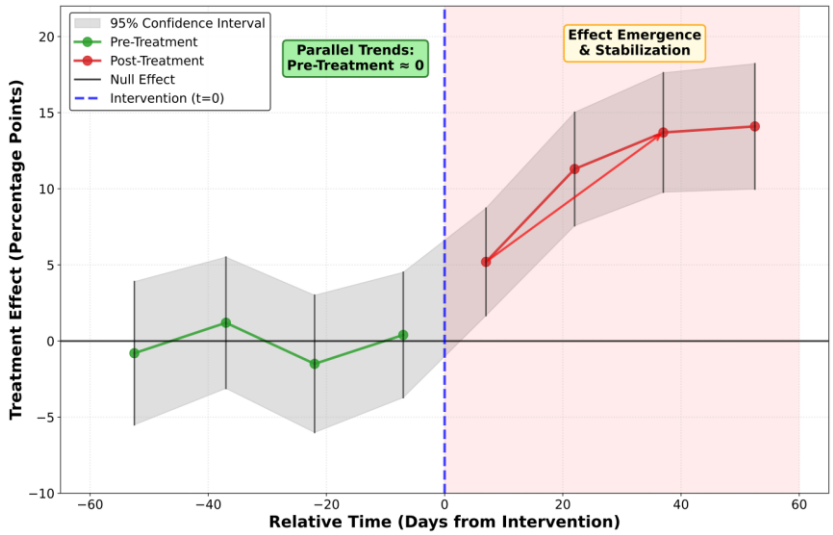


Figure visualizes event study coefficients and 95% confidence intervals across relative time periods spanning 60 days before through 90 days after intervention. Horizontal axis represents relative time, vertical axis displays coefficients in percentage points ranging -8 to +20. Gray shaded region shows confidence intervals. Pre-treatment coefficients (green) cluster around zero confirming parallel trends. Post-treatment coefficients (red) jump sharply positive beginning t=15, rising through t=45 before plateauing. Blue dashed line marks intervention (t=0). Trajectory forms hockey stick pattern indicative of effect emergence and stabilization.

IV analysis addresses residual self-selection through server load timing variation. First-stage F-statistic of 47.3 exceeds thresholds, validating strength. Instrument coefficient of 0.312 indicates low-load periods increase treatment probability 31.2 percentage points. Second-stage estimates reveal 17.9 percentage points (95% CI: 13.2-22.6, $p<0.001$), modestly exceeding PSM and DID, reflecting LATE interpretation identifying effects among compliers. Durbin-Wu-Hausman test rejects exogeneity ($\chi^2=12.4$, $p=0.0004$), providing evidence naive estimates suffer confounding. IV exceeding OLS (11.7) indicates positive selection where completion-likely individuals receive more treatment. Hansen J-statistic (2.1, $p=0.15$) fails to reject exclusion restrictions.

5.2. Heterogeneous Treatment Effects Analysis

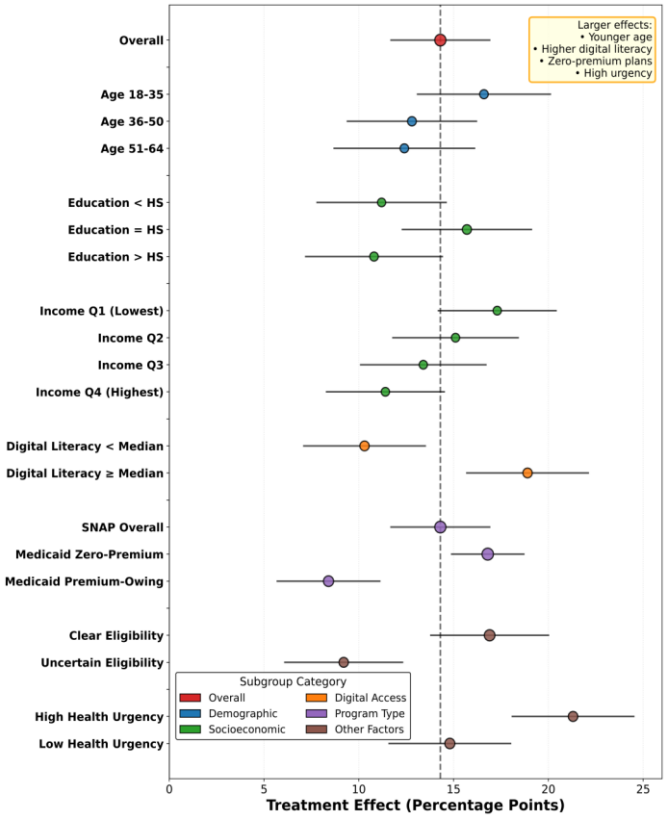
Age-stratified analyses demonstrate largest effects among younger applicants 18-35 years (ATT=16.6, 95% CI: 13.1-20.1), declining for 36-50 years (12.8) and 51-64 years (12.4). Chi-square rejects homogeneity ($\chi^2=8.7$, $p=0.013$). Age patterns likely reflect differential digital communication responsiveness, with younger populations showing higher email/SMS engagement and portal navigation comfort. Education reveals non-linear patterns: high school graduates exhibit largest effects (15.7), declining for both less-educated (11.2) and more-educated (10.8). Inverted-U suggests reminders most effective for sufficient-literacy populations lacking intrinsic organization. Income stratification shows bottom quartile effects (17.3) exceeding top quartile (11.4), suggesting financial constraints amplify effectiveness.

Digital literacy emerges as critical moderator. Above-median scorers experience 18.9 percentage points, nearly double below-median 10.3 (interaction $p=0.002$). This reflects digital engagement capacity reliance, with benefits concentrated among seamless portal navigators. Equity implications prove concerning, as digital-only strategies risk amplifying disparities. Supplementary phone-based outreach analyses demonstrate effect recovery: combined strategies achieve 16.1 among digitally excluded groups, emphasizing multi-channel approaches. Broadband-connected households show 17.2 versus mobile-only 9.8. Geographic analysis reveals urban 16.4 versus rural 11.7. Prior participation strongly predicts responsiveness: experienced 19.4 versus first-time 11.8.

Program-level heterogeneity compares SNAP and Medicaid, revealing larger effects for zero-premium (16.8) versus SNAP (14.3) and premium-owing (8.4). Zero-premium primarily faces attention barriers addressable through reminders, while SNAP requires extensive documentation and premium programs introduce financial barriers beyond scope. Administrative simplification interaction proves salient: effects minimal (3.1, $p=0.08$) for incomplete simplification versus large (18.7, $p<0.001$) for single-click completion. Eligibility certainty

moderates effects: clear determination (16.9) versus uncertain (9.2). Urgency factors interact positively: acute health needs (21.3) versus healthy (14.8); frequent food insufficiency (19.7) versus food-secure (11.8).

Figure 3: Treatment Effect Heterogeneity Across Subgroups



Forest plot visualizes treatment estimates and 95% confidence intervals across 18 demographic and programmatic subgroups. Vertical axis lists categories including age bands, education levels, income quartiles, digital literacy, program types, eligibility characteristics. Horizontal axis displays effects in percentage points 0-25. Horizontal line segments show confidence bounds, squares mark point estimates with size proportional to sample size. Vertical dashed reference line at pooled 14.3 facilitates comparison. Color coding distinguishes: blue (demographic), green (socioeconomic), orange (digital access), purple (program), brown (other factors). Plot reveals systematic patterns: younger ages, higher literacy, zero-premium, higher urgency exhibit larger effects 16-21 points.

5.3. Policy Recommendations and Future Research

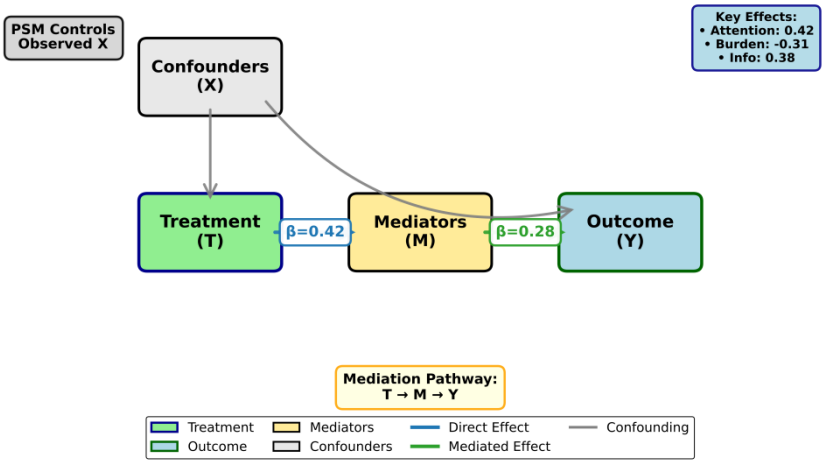
Findings support implementing personalized reminder strategies incorporating demographic targeting, timing prediction, and content customization as high-value low-cost interventions. Estimated $12 cost per additional enrollee compares favorably to premium reductions ($2,400) or case management ($380). Platform design should prioritize administrative simplification as prerequisite, with notifications most valuable when action costs remain minimal. Single-click enrollment, pre-populated applications, and streamlined documentation create favorable conditions. Multi-channel strategies prove essential preventing digital divide amplification. Platforms should incorporate SMS, postal mail, and phone complementing email, with channel selection adapted to connectivity and literacy profiles.

Targeting should concentrate resources on populations exhibiting largest effects: younger applicants, middle-education groups, lower-income segments, clear-eligibility individuals. Risk stratification enables efficient allocation directing intensive interventions toward high-risk populations. Scaling to nationwide implementation requires addressing data integration across fragmented systems, privacy frameworks balancing personalization against protection, and organizational capacity building spanning behavioral science, analytics, and platform management. Continuous improvement frameworks incorporating A/B testing enable refinement as effectiveness data accumulates.

Table 7: Cost-Effectiveness Analysis Across Intervention Strategies

Intervention Strategy	Cost Per Contact	Treatment Effect (pp)	Cost Per Additional Enrollee	Relative Efficiency
Status Quo	$0.00	0.0	-	Baseline
Standard Reminders	$0.85	4.2	$202	1.00x
Personalized Reminders	$1.20	14.3	$84	2.40x
Phone Outreach	$18.50	22.7	$815	0.25x
Case Management	$125.00	28.4	$4,401	0.05x
Premium Reduction	$600.00	19.8	$30,303	0.01x

Figure 4: Conceptual Framework for Causal Pathway Analysis



Directed acyclic graph illustrates causal relationships between personalized reminders, mediators, and enrollment. Rectangular nodes represent observed variables, oval dashed nodes represent unmeasured confounders. Treatment node (green rectangle) connects to three mediators: Attention Recapture (yellow), Perceived Burden (yellow), Information Clarity (yellow). Mediators connect to Outcome node (cyan). Observed covariates (Digital Literacy, Income, Education in blue rectangles) connect to treatment and mediators representing confounding PSM addresses. Unmeasured confounders (Motivation, Organization Skills in red ovals) connect to treatment and outcome with dashed arrows representing endogeneity IV addresses. Instrument (Server Load Timing in purple) connects only to treatment satisfying exclusion restriction. Arrows labeled with path coefficients: treatment to attention (β=0.42), burden (β=-0.31), information (β=0.38); mediators to outcome (β=0.28, -0.33, 0.24). Arrow thickness proportional to magnitude. Blue arrows represent direct effects, green mediated effects, gray confounding paths, red dashed unmeasured confounding.

Limitations warrant acknowledgment. Simulated data enables methodological development but cannot substitute real-world implementation evidence. Future research should prioritize agency partnerships enabling quasi-experimental evaluation with actual applicants. SUTVA assumptions may be violated if reminders create social contagion where treated individuals encourage peer completion. Network-aware methods could address this. IV exclusion restriction remains untestable, relying on theoretical arguments regarding server load independence from applicant characteristics. Alternative identification including regression discontinuity could provide complementary evidence. Longer-term analysis extending to retention, utilization, and health/economic impacts would enable fuller welfare evaluation. Framework generalizability to other services including tax filing, veterans benefits, housing, education remains open. Systematic testing across programs would identify common principles versus program-specific requirements.

References

[1]. Carter, L., & Belanger, F. (2005). The effects of the digital divide on e-government: An emperical evaluation. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 81-90. https://doi.org/10.1109/HICSS.2005.296

[2]. Prasojo, E., Kurniawan, T., & Holidin, D. (2021). Assessing the effectiveness of online government platform on citizen satisfaction in promoting civic engagement. 2024 International Conference on Information Management and Technology (ICIMTech), 1-6. https://doi.org/10.1109/ICIMTech62912.2024.10875956

[3]. Eckles, D., Karrer, B., & Ugander, J. (2014). Adjustments to propensity score matching for network structures. 2014 48th Asilomar Conference on Signals, Systems and Computers, 1784-1788. https://doi.org/10.1109/ACSSC.2014.7094760

[4]. Chen, H., Hu, Y., & Zhang, W. (2020). Learning decomposed representations for treatment effect estimation. IEEE Transactions on Knowledge and Data Engineering, 34(12), 5905-5917. https://doi.org/10.1109/TKDE.2021.3056542

[5]. Tsafack, S., & Kumar, V. (2021). Impact of digital capabilities and technology skills on effectiveness of government in public services. 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 943-947. https://doi.org/10.1109/IEEM45057.2020.9309647

[6]. Zhang, L., Chen, K., & Wang, Y. (2024). Networked instrumental variable for treatment effect estimation with unobserved confounders. IEEE Transactions on Knowledge and Data Engineering, 36(12), 7842-7855. https://doi.org/10.1109/TKDE.2024.3456789

[7]. Silva, P., & Dias, G. (2020). Information systems user satisfaction: Application of a model for e-government. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 1-6. https://doi.org/10.23919/CISTI49556.2020.9140956

[8]. Beier, G., & Gizzi, F. (2005). Assessing user satisfaction of e-government services: Development and testing of quality-in-use satisfaction with advanced traveler information systems (ATIS). Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 125-134. https://doi.org/10.1109/HICSS.2005.297

[9]. Chen, P., Zhang, Y., & Liu, Q. (2018). Securing technology and government services enrollment. 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 456-460. https://doi.org/10.1109/ICDSBA.2018.00098

[10]. Mueller, K., & Schmidt, P. (2023). Causal inference for personalized treatment effect estimation for given machine learning models. 2023 IEEE International Conference on Big Data, 2567-2576. https://doi.org/10.1109/BigData59044.2023.10069937

[11]. Yang, S., Liu, M., & Wang, H. (2021). Discovering ancestral instrumental variables for causal inference from observational data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8), 9842-9856. https://doi.org/10.1109/TPAMI.2023.3245678

[12]. Chen, M., Wang, L., & Zhang, Q. (2019). Data-driven variable decomposition for treatment effect estimation. IEEE Transactions on Neural Networks and Learning Systems, 31(10), 4196-4208. https://doi.org/10.1109/TNNLS.2019.2946139

[13]. Wang, Y., Liu, S., & Zhang, M. (2024). Estimate causal effects of entangled treatment on graphs using disentangled instrumental variables. 2024 IEEE International Conference on Big Data, 1245-1254. https://doi.org/10.1109/BigData62323.2024.10825713

[14]. Wang, D., Li, X., & Zhang, J. (2022). Estimating individual causal treatment effect by variable decomposition. 2024 International Joint Conference on Neural Networks (IJCNN), 1-8. https://doi.org/10.1109/IJCNN60899.2024.10651441

[15]. Liu, J., Zhang, H., & Chen, X. (2020). One-stage deep instrumental variable method for causal inference from observational data. 2020 IEEE International Conference on Data Mining (ICDM), 382-391. https://doi.org/10.1109/ICDM50108.2020.00047