Robust Evaluation of Customer Personalization Algorithms in Healthcare Using Causal Inference and Offline Policy Testing

Santiago Rojas[1], Felipe Cárdenas[2], Mauricio Villalba[3]

[1]Department of Systems Engineering, Universidad del Cauca, Calle 5 No. 4-70, Popayán, Colombia
[2]Department of Informatics and Telecommunications, Universidad de Córdoba, Carrera 6 No. 76-103, Montería, Colombia
[3]Department of Computer and Information Sciences, Universidad de Nariño, Calle 18 Carrera 50, Ciudadela Universitaria Torobajo, Pasto, Colombia

ABSTRACT Because clinical environments are safety-critical and heavily constrained by privacy, regulation, and limited experimentation capacity, many organizations evaluate personalization strategies using historical observational data rather than prospective randomized trials. This paper develops a technical framework for robust evaluation of customer personalization algorithms in healthcare using causal inference and offline policy testing. We formalize personalization as a policy that maps patient context to actions such as messaging intensity, appointment scheduling options, or care-management enrollment. We analyze identifiability under potential outcomes, clarify assumptions required for counterfactual generalization, and highlight how standard predictive validation can systematically misestimate patient benefit under selection and feedback. We then derive offline policy value estimators based on importance sampling, direct modeling, and doubly robust constructions, emphasizing variance control, positivity diagnostics, cross-fitting, and uncertainty quantification. Robustness is treated as a first-class objective through sensitivity analysis for unmeasured confounding, distribution shift between logging and deployment populations, and outcome missingness mechanisms common in longitudinal healthcare data. Finally, we translate these methods into an operational evaluation workflow that supports governance, fairness auditing, and safety constraints without relying on live experimentation. The resulting approach supports comparative testing of candidate personalization algorithms while explicitly characterizing the causal and statistical conditions under which offline estimates are reliable.

I. Introduction

Personalization in healthcare refers to algorithmic mechanisms that adapt decisions to patient-level context [1]. In this paper, the term customer is used in its operational sense common to patient engagement platforms, where individuals interact with digital, administrative, or clinical services and where system operators seek to tailor interventions to improve outcomes such as appointment adherence, medication persistence, timely follow-up, preventive screening uptake, or chronic disease monitoring. Unlike recommendation settings where the primary objective is often short-term engagement, healthcare personalization is constrained by safety requirements, complex outcome measurement, and ethical obligations to avoid disproportionate harm. Moreover, healthcare outcomes are frequently delayed, partially observed, and confounded by clinical decision processes, insurance coverage, and patient preferences. These characteristics make naive offline evaluation unreliable when it relies solely on predictive accuracy or on comparisons of observed outcomes across algorithmic segments.

Offline evaluation is attractive because it can be performed without exposing patients to experimental policies, and because many healthcare systems maintain extensive logs of actions and outcomes [2]. However, these logs are typically generated by historical decision policies that were not randomized. Clinicians and operational teams decide who receives which action using rules and judgments that incorporate latent information, and patients self-select into care pathways. Consequently, the observed relationship between an action and an outcome is often biased with respect to the causal effect of that action. A personalization algorithm that appears to perform well under standard supervised metrics may simply mimic existing triage patterns, allocating higher-intensity outreach to patients already likely to improve or to those with more complete follow-up data. Conversely, an algorithm that would meaningfully improve outcomes could

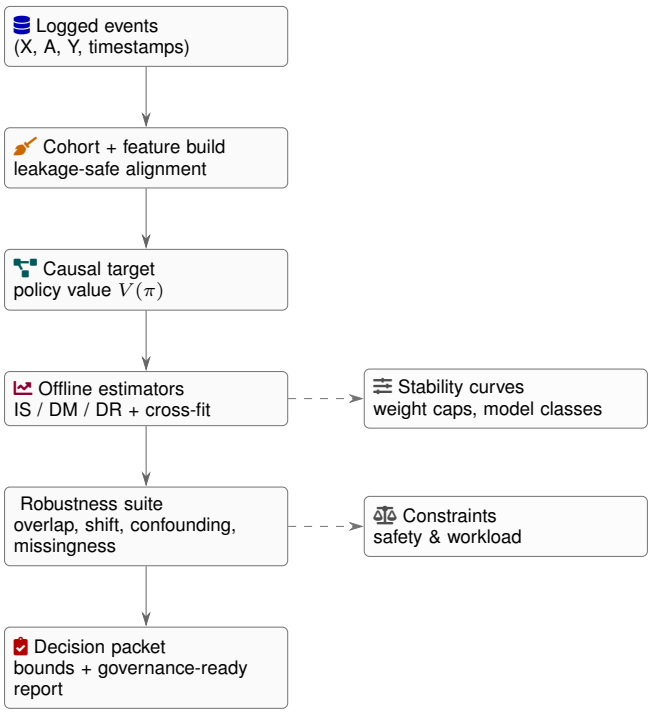

FIGURE 1: End-to-end offline evaluation pipeline for healthcare personalization: from logged decision instances to causal policy-value estimates, robustness diagnostics, and a governance-ready deployment packet.

look poor in logs if it targets patients whose outcomes are historically worse and whose engagement is more difficult.

Causal inference offers a principled language for translating logged observational data into counterfactual statements about what would have happened under alternative policies [3]. Yet applying causal inference to personalization requires careful attention to the structure of decision-making, the timing of information, and the repeated nature of interventions. Even when causal identification is plausible, the evaluation problem differs from estimating a single average treatment effect. A personalization policy may choose different actions for different contexts, and its value depends on the joint distribution of contexts, actions, and potential outcomes. Offline policy testing, also called off-policy evaluation, aims to estimate the expected outcome that would result if a candidate policy were deployed, using data collected under a different logging policy. This setting is common in contextual bandits and reinforcement learning, but healthcare introduces additional complications such as censoring, dynamic confounding, and safety constraints that require the evaluation to be robust rather than merely point-estimable.

Robust evaluation in this context has multiple meanings [4]. Statistical robustness concerns estimators that remain stable under model misspecification, high variance from rare actions, and finite-sample uncertainty. Causal robustness concerns sensitivity to violations of ignorability, including unmeasured severity, unrecorded social determinants, and clinician intuition. Operational robustness concerns distribution shift, for example when a new policy changes patient engagement patterns or when the patient population changes due to seasonality, payer contracts, or epidemics. Regulatory robustness concerns explainability, auditing, and reproducibility, including the ability to justify assumptions and to document monitoring plans. The methods in this paper are organized around these interlocking notions of robustness, with an emphasis on mechanisms that allow a healthcare organization to quantify uncertainty and to detect when offline estimates are not reliable enough to support deployment decisions.

We consider personalization algorithms that map a patient context vector to an action [5]. Context can include demographics, prior utilization, lab results, diagnoses, risk scores, behavioral signals, and operational constraints. Actions can include sending an SMS reminder, calling a patient, offering telehealth versus in-person scheduling, enrolling the patient in a care-management program, escalating to a nurse, adjusting follow-up frequency, or selecting content variants for education. Outcomes can include binary endpoints such as completion of a visit, time-to-event outcomes such as time to readmission, or continuous measures such as HbA1c levels. In many operational settings the algorithm is evaluated against multiple objectives, including clinical benefit, patient burden, and cost. We focus on a generic scalar outcome that can encode trade-offs through a utility

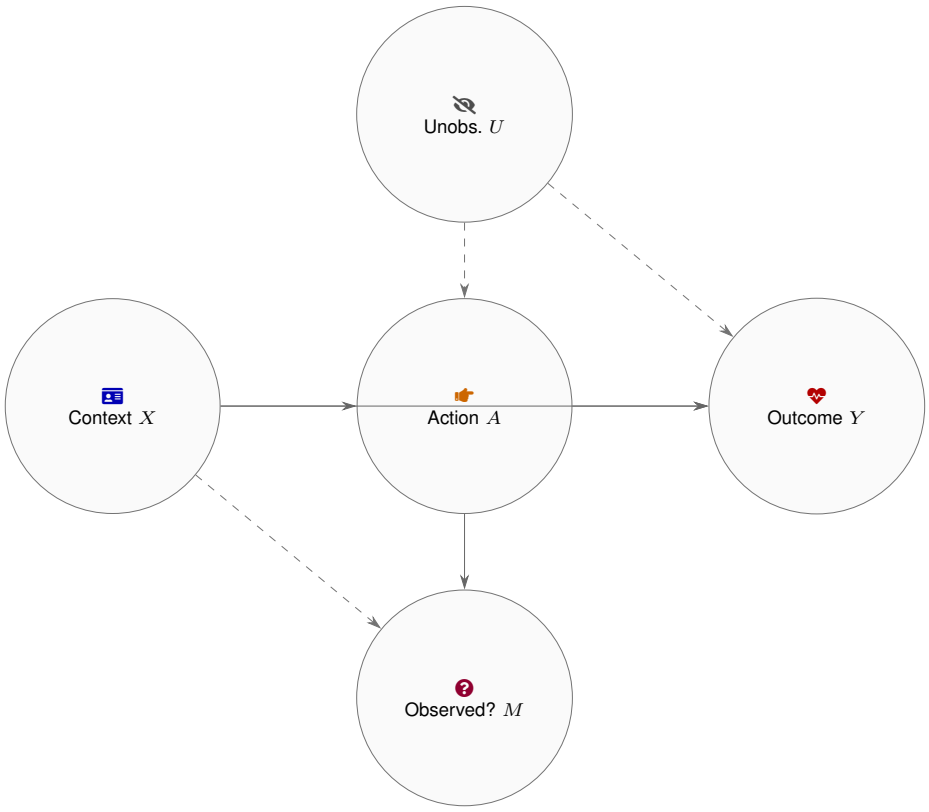

FIGURE 2: Causal structure for offline policy testing. Identifiability of policy value relies on conditioning on context X to address confounding; dashed edges highlight unmeasured confounding (U) and outcome missingness via an observation indicator (M).

function, while noting that multi-objective extensions follow by vector-valued outcomes and constraint handling.

The core contribution of the paper is a coherent, end-to-end evaluation framework [6]. We define a data-generating process that captures logged actions and outcomes with potentially time-varying covariates. We specify the causal estimand as the value of a candidate policy and relate it to the g-formula under standard assumptions. We then present offline policy estimators, including importance sampling, direct outcome regression, and doubly robust estimators with cross-fitting, emphasizing practical variance control and positivity diagnostics. We further propose robustness analyses that quantify sensitivity to unmeasured confounding and to distribution shift, and we discuss uncertainty quantification that yields conservative intervals suitable for safety-critical decisions. Finally, we translate these methods into an operational workflow that integrates privacy, governance, and fairness considerations, acknowledging that healthcare personalization is not merely a statistical exercise but a socio-technical system embedded in clinical operations.

The remainder of the paper assumes familiarity with basic probability and statistical learning but does not require specialized reinforcement learning background [7]. Mathematical expressions are provided where they clarify the structure of estimators and assumptions. Throughout, notation is chosen to emphasize the distinction between observed variables and potential outcomes, and to make explicit the role of the logging policy that generated the data. The goal is not to advocate a single estimator but to provide a robust evaluation toolbox with explicit diagnostics and failure modes, allowing practitioners to select methods that match the reliability requirements of a given deployment decision.

II. Problem Setting and Data Generating Process

Let X denote a patient context vector observed at the time a decision is made. Let A denote the action selected by the operational system, where A takes values in a finite set $\mathcal{A}$ or a bounded continuous set. Let Y denote an outcome measured after action assignment, potentially after a delay. Logged healthcare data often contains additional variables that are not purely pre-treatment, such as intermediate measures, follow-up actions, or clinician notes recorded after the initial action [8]. For offline policy testing, it is essential to define a decision point at which X is available, A is chosen, and Y is evaluated, while avoiding leakage of post-action information into X. In practice, constructing X requires

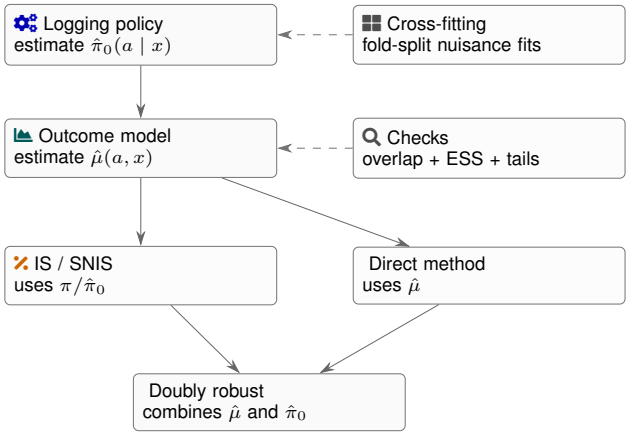

FIGURE 3: Estimator stack for offline policy evaluation. Importance sampling reweights by $\pi/\hat{\pi}_0$, direct methods plug in $\hat{\mu}$, and doubly robust estimators combine both with cross-fitting and overlap diagnostics to improve finite-sample reliability.

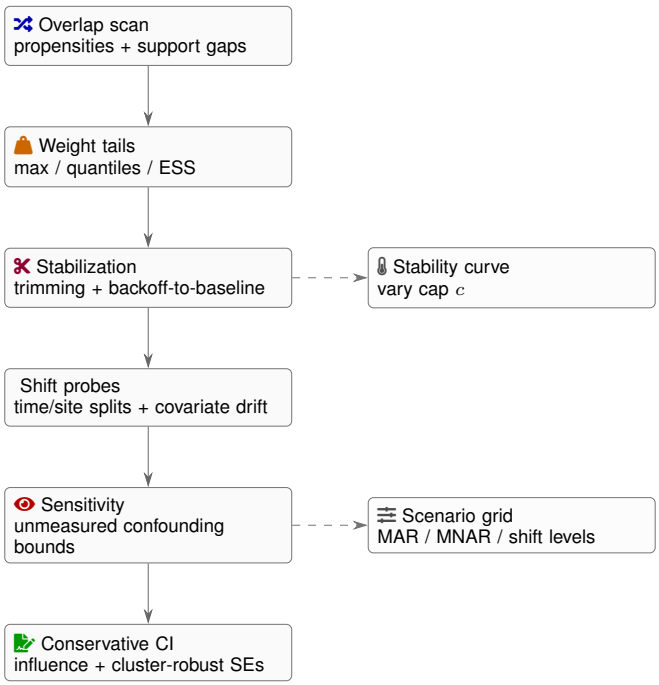

FIGURE 4: Robustness diagnostics workflow. Evaluation is accepted only when overlap is adequate, weight tails are controlled, results are stable to trimming/backoff, and conclusions persist under plausible distribution-shift and confounding sensitivity scenarios.

careful timestamp alignment to ensure that all covariates precede the action.

We assume the logged dataset consists of n independent patient-decision instances indexed by $i \in \{1, \ldots, n\}$, each containing (X_i, A_i, Y_i). Independence can be an approximation when decisions are spaced sufficiently apart or when cluster dependence is weak. However, healthcare logs often contain repeated decisions per patient and correlated outcomes due to shared clinicians or facilities. A more faithful model uses clustered or longitudinal data, but many offline policy estimators can be adapted through cluster-robust variance estimation and through trajectory-based formulations [9]. We begin with the single-decision setting to present core ideas and then discuss extensions.

A key object is the logging policy, denoted $\pi_0(a \mid x)$, which describes the probability that the historical system chooses action a given context x. In healthcare, π_0 may be partly deterministic due to rules, and partly stochastic due to clinician variability, scheduling availability, and unrecorded factors. Offline evaluation requires either explicit knowledge of π_0 or a reliable estimate of it from logs. If the logging policy is unknown and the data is deterministically logged,

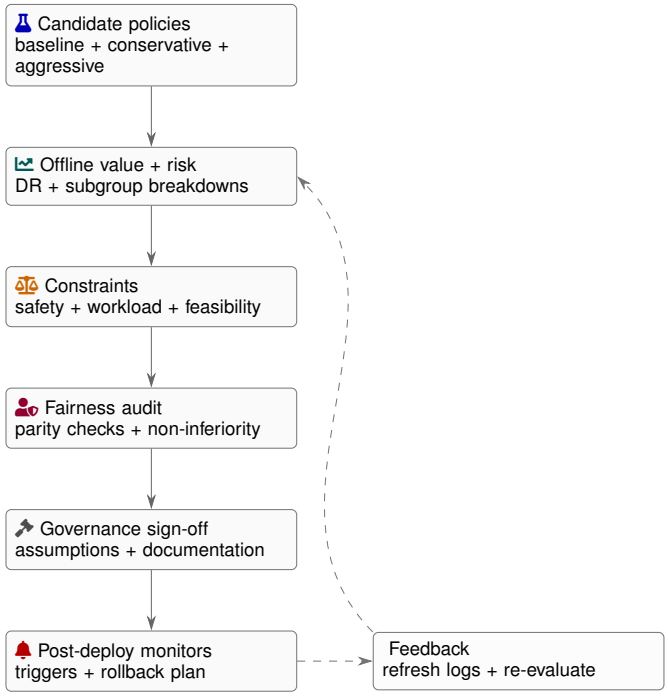

FIGURE 5: Operational governance loop. Offline causal evidence, constraints, and fairness auditing feed a sign-off process, while monitoring and feedback ensure the system remains safe under drift and policy-induced changes.

the evaluation problem becomes ill-posed for actions that were rarely or never taken in some contexts. This connects to the positivity condition, discussed later [10].

We define a candidate personalization algorithm as a policy $\pi(a \mid x)$, which may be deterministic, stochastic, or parameterized by a model. Deterministic policies select $a = \pi(x)$, while stochastic policies allow randomized exploration or risk-aware diversification. In safety-critical healthcare contexts, stochasticity is sometimes constrained, but it can still be used in limited, controlled ways, for example by randomizing among clinically equivalent message templates or among scheduling options that preserve clinical safety.

Outcomes in healthcare require careful construction. Suppose Y is a scalar utility that increases with desirable clinical events and decreases with cost or burden. For example, Y could be defined as a weighted combination of appointment completion, medication refill adherence, and patient-reported satisfaction, where weights are chosen by clinical governance [11]. If outcomes are delayed or censored, one can define Y as a function of observed follow-up time, such as a restricted mean survival time up to a horizon. If outcomes are missing due to loss to follow-up, the evaluation must model the missingness mechanism or incorporate inverse probability weights for observation.

In many personalization systems, decisions occur repeatedly. Let $t \in \{1, \ldots, T\}$ index decision times, and let H_t denote the history available at time t, including baseline context and prior actions and observations. Let A_t denote the action at time t, and let Y denote a terminal outcome, or let R_t denote rewards at each time. The data is then a trajectory $(H_1, A_1, R_1, \ldots, H_T, A_T, R_T)$. Offline policy testing in this setting becomes off-policy evaluation in a Markov or partially observable process, with additional complications from time-varying confounding [12]. Although trajectory evaluation is important in care management, many healthcare engagement decisions can be approximated as episodic single-decision problems by choosing a specific decision point, such as the first outreach attempt for a missed appointment or the selection of a post-discharge follow-up plan. The framework in this paper supports both, but the mathematical exposition emphasizes the single-decision case for clarity.

We introduce the potential outcomes notation. Let $Y(a)$ denote the potential outcome that would be observed if action a were taken, possibly contrary to the action actually taken. Under a policy π, the potential outcome is $Y(\pi)$, which can be defined as $Y(\pi) = \sum_{a \in \mathcal{A}} Y(a) \mathbb{I}\{A^\pi = a\}$ where A^π is the action assigned by π given X. The causal estimand of interest is the policy value, defined as the expected potential outcome under the policy. In the single-decision case this is [13]

$$\begin{aligned} V(\pi) &= \mathbb{E}\big[Y(\pi)\big] \\ &= \mathbb{E}\Big[\sum_{a \in \mathcal{A}} Y(a)\, \pi(a \mid X)\Big] \end{aligned}$$

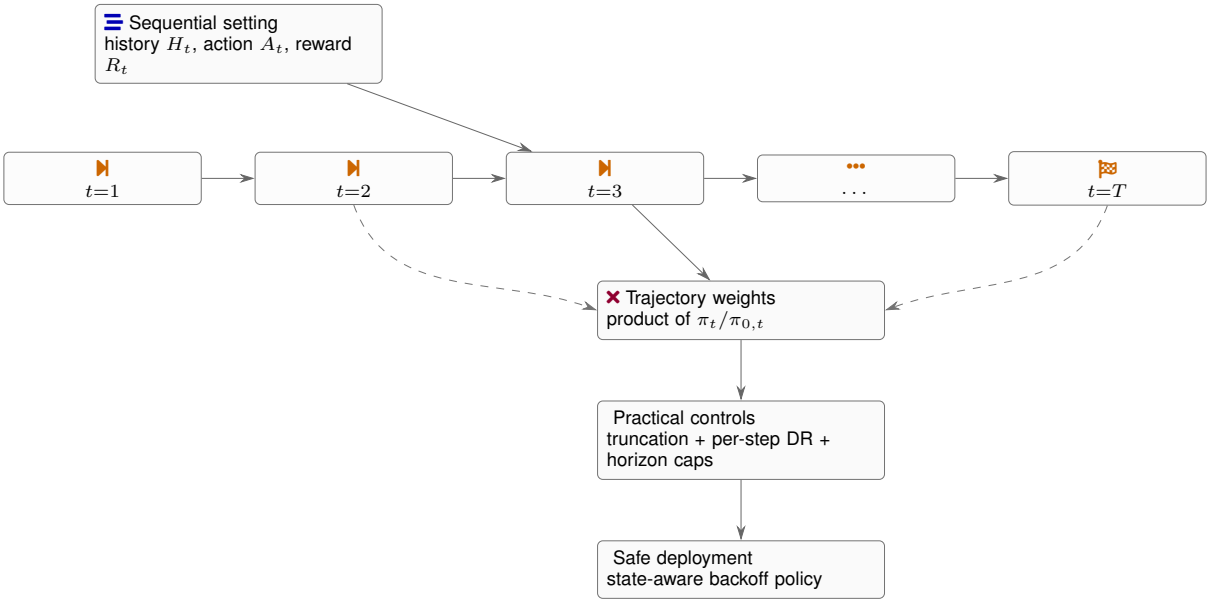

FIGURE 6: Extension to repeated decisions. Sequential off-policy evaluation multiplies per-time-step ratios, which can become unstable; healthcare deployments typically enforce truncation, per-step doubly robust corrections, and state-aware backoff rules for safety.

where the expectation is over the distribution of X and the potential outcomes. In a deterministic policy, $\pi(a \mid X)$ is an indicator that selects one action.

The evaluation challenge is that we only observe $Y_i = Y_i(A_i)$ for the action actually taken in the logs [14]. Recovering $V(\pi)$ from logged data requires assumptions and estimators that combine information across contexts and actions. Importantly, the evaluation target may be either in-sample, reflecting the historical distribution of contexts, or out-of-sample, reflecting a target population. In healthcare, distribution shift can occur due to changes in enrollment, new clinics, seasonal illness, or operational policy changes. Therefore, it is often necessary to define a target context distribution $p^\star(x)$ and to evaluate policy value under that distribution. When $p^\star(x)$ differs from the logged distribution $p(x)$, evaluation requires additional weighting or transportability assumptions. A robust evaluation framework should explicitly state whether the value is defined for the logged population, a future population, or a mixture, because interpretability and risk differ across these choices [15].

Data quality is central. Logged healthcare data often contains measurement error, coding drift, missingness, and delayed outcome capture. These issues affect both identification and estimation. For example, diagnosis codes may be incomplete, and social determinants may be absent, leading to unmeasured confounding. Moreover, the action variable may be imperfectly logged, such as when outreach attempts are not fully recorded or when patient exposure is uncertain [16]. Robust evaluation must treat these as potential failure points and incorporate diagnostics, sensitivity analysis, and conservative uncertainty.

III. Causal Inference Foundations for Personalization

The central causal question is whether we can express the policy value $V(\pi)$ in terms of observable quantities from the logged data. A common identification approach relies on three conditions. The first is consistency, which states that the observed outcome equals the potential outcome corresponding to the observed action, meaning $Y = Y(A)$. This requires that the action variable captures the relevant intervention and that there are no multiple versions of treatment that differ in ways not encoded by A. In healthcare, this can fail if the action category aggregates heterogeneous implementations, such as phone calls of different quality or care-management enrollment with varying intensity [17]. A robust evaluation pipeline should therefore refine the action representation or incorporate additional variables that capture relevant variation.

The second condition is conditional exchangeability, also called ignorability, which states that given the observed context X, the action assignment is as good as random with respect to potential outcomes. Formally, for all a,

$$Y(a) \perp A \mid X$$

This assumption is strong in healthcare because clinicians may use latent severity and patient motivation not captured in X. However, in certain operational actions, such as randomized message templates or scheduling options constrained by availability, ignorability can be plausible after conditioning

on rich administrative covariates [18]. When ignorability is questionable, sensitivity analysis becomes necessary, and the evaluation should be interpreted as conditional on the assumption rather than as definitive causal truth.

The third condition is positivity, which requires that the logging policy assigns nonzero probability to actions that the candidate policy might choose, for contexts of interest. For discrete actions,

$$\pi_0(a \mid x) > 0 \quad \text{whenever} \quad \pi(a \mid x) > 0$$

In healthcare, positivity violations are common because some actions are clinically contraindicated in certain contexts or because operational constraints prevent certain options [19]. From an evaluation standpoint, positivity violations mean that we cannot learn the counterfactual outcome for those actions in those contexts from logged data alone. Robust evaluation thus requires either restricting the candidate policy to the support of the logging policy, aggregating contexts to improve overlap, or incorporating external evidence such as clinical trials, which is beyond the scope of purely offline evaluation.

Under consistency, exchangeability, and positivity, the policy value is identified by the g-formula:

$$V(\pi) = \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \mathbb{E}\big[Y \mid X, A = a\big]\, \pi(a \mid X)\Big] \text{[20]}$$

This expression clarifies that evaluation can proceed by modeling the conditional outcome regression $\mu(a, x) = \mathbb{E}[Y \mid A = a, X = x]$ and then averaging it under the candidate policy. However, pure outcome regression can be sensitive to model misspecification, especially in high-dimensional healthcare contexts with nonlinear relationships and interactions.

A causal perspective also highlights why standard predictive validation is insufficient. A supervised model that predicts Y from (X, A) might achieve high accuracy without correctly estimating counterfactual differences between actions. For instance, if A is strongly correlated with unobserved severity, the model may attribute poor outcomes to the action rather than to severity. Even if the model predicts well, its implied counterfactuals may be biased. This motivates estimators designed for causal targets and robust to certain model errors [21].

The personalization objective often involves heterogeneity. A useful causal quantity is the conditional average treatment effect, defined for two actions a and a' as

$$\tau_{a,a'}(x) = \mathbb{E}\big[Y(a) - Y(a') \mid X = x\big]$$

A policy can be viewed as choosing the action that maximizes expected outcome given x, subject to constraints. Yet offline evaluation focuses on comparing fixed candidate policies rather than on learning the optimal one, because robust evaluation and robust optimization are distinct tasks. Nevertheless, understanding heterogeneity is important for diagnosing where a candidate policy might help or harm, and for defining safety constraints such as requiring non-inferiority in certain subgroups.

Healthcare personalization must also contend with interference and spillovers [22]. For example, outreach capacity is limited; prioritizing one patient may delay another. This violates the stable unit treatment value assumption because one patient's outcome depends on others' assigned actions. Offline evaluation under interference requires modeling the joint assignment and capacity constraints, often using cluster-level policies or treating capacity as part of the state. In many practical engagement settings, interference is moderate but not negligible, especially when actions consume scarce resources like nurse calls. A robust evaluation should therefore either restrict to actions with minimal interference, incorporate capacity as a covariate, or interpret results as conditional on approximate non-interference [23].

Another fundamental issue is time. Even in a single-decision framing, the outcome may depend on post-action events such as subsequent clinician decisions that are influenced by the action. For example, a reminder might increase visit completion, which then triggers additional diagnostics, affecting downstream outcomes. Depending on the outcome definition, such downstream pathways may be desired and part of the causal effect, or they may be considered mediators. Offline policy testing typically targets the total effect of assigning actions under existing downstream processes. This is appropriate for operational decisions where the downstream system remains unchanged [24]. If deployment changes downstream processes, such as altering clinician workload and thus follow-up decisions, then the offline estimate may not transport well. This links to robustness under policy-induced distribution shift, discussed later.

Because healthcare data is often longitudinal, dynamic treatment regimes provide a formalism for sequential personalization. Let π map histories to actions. The value becomes an expectation over trajectories, and identification requires sequential ignorability, meaning that at each time, given history, the action is independent of future potential outcomes. In practice, sequential ignorability is often less plausible than single-time ignorability because unmeasured factors accumulate [25]. Therefore, robust offline evaluation in longitudinal settings frequently relies on rich logging and careful representation of history, combined with conservative sensitivity analysis.

IV. Offline Policy Testing and Counterfactual Risk

Offline policy testing aims to estimate $V(\pi)$ using data generated under π_0. A baseline estimator uses inverse propensity weighting, also called importance sampling. Define the

weight

$$w(X, A) = \frac{\pi(A \mid X)}{\pi_0(A \mid X)}$$

Assuming π_0 is known or estimated, an importance sampling estimator is [26]

$$\widehat{V}_{\mathrm{IS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} w(X_i, A_i)\, Y_i$$

This estimator is unbiased under the identification assumptions when π_0 is correct, but it can have very high variance if π chooses actions that were rare under π_0 in some regions of context space. In healthcare, such rare actions can arise when candidate policies propose increased outreach to historically under-served patients, leading to large weights and unstable estimates. Variance control is thus essential.

A common variance-reduced variant is self-normalized importance sampling: [27]

$$\widehat{V}_{\mathrm{SNIS}}(\pi) = \frac{\sum_{i=1}^{n} w(X_i, A_i)\, Y_i}{\sum_{i=1}^{n} w(X_i, A_i)}$$

Self-normalization introduces bias but often reduces variance substantially. In safety-critical contexts, it is important to quantify both bias and variance and to avoid relying on a single estimator. Truncation or clipping of weights is another approach, where w is replaced by $\min(w, c)$ for some threshold c [28]. This reduces variance at the cost of bias, and robust evaluation requires exploring the bias-variance trade-off and interpreting results conservatively.

Direct methods estimate the outcome regression $\mu(a, x)$ and then compute

$$\widehat{V}_{\mathrm{DM}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \widehat{\mu}(a, X_i)\, \pi(a \mid X_i)$$

Direct methods can be low-variance but can be biased under model misspecification. Healthcare outcomes can be nonlinear and heteroskedastic, and outcome models may be mis-specified even with flexible learners due to limited overlap and due to systematic missingness.

Doubly robust estimators combine both propensity weighting and outcome regression [29]. One canonical form is

$$\widehat{V}_{\mathrm{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{a \in \mathcal{A}} \widehat{\mu}(a, X_i)\, \pi(a \mid X_i) \right] + \frac{1}{n} \sum_{i=1}^{n} w(X_i, A_i)\, [30] \Big[Y_i - \widehat{\mu}(A_i, X_i) \Big]$$

The doubly robust property states that, under regularity conditions, the estimator is consistent if either the propensity model or the outcome model is correctly specified. In practice, both are estimated, often with flexible machine learning models, and finite-sample performance depends on estimation error and overlap. Cross-fitting, where models are trained on folds and evaluated on held-out folds, helps reduce overfitting bias and improves asymptotic properties in high-dimensional settings.

To express cross-fitting compactly, let $\mathcal{I}_k$ be indices in fold k, and let $\widehat{\mu}^{(-k)}$ and $\widehat{\pi}_0^{(-k)}$ be models trained excluding fold k. Then

$$\widehat{V}_{\mathrm{DR,CF}}(\pi) = \frac{1}{n} \sum_{k} \sum_{i \in \mathcal{I}_k} \left[[31] \sum_{a \in \mathcal{A}} \widehat{\mu}^{(-k)}(a, X_i)\, \pi(a \mid X_i) \right] + \frac{1}{n} \sum_{k} \sum_{i \in \mathcal{I}_k} \frac{\pi(A_i \mid X_i)}{\widehat{\pi}_0^{(-k)}(A_i \mid X_i)} \Big[Y_i - \widehat{\mu}^{(-k)}(A_i, X_i) \Big]$$

Cross-fitting is particularly valuable in healthcare where X can include thousands of sparse diagnosis features and where propensity estimation may overfit.

When actions are continuous, importance sampling can be unstable because densities can be extremely small [32]. One approach is to restrict evaluation to policies that stay close to the logging policy, such as Gaussian perturbations around clinician defaults, or to discretize actions into clinically meaningful bins. Another approach uses kernel-based density ratio estimation, but this requires strong smoothness and overlap. In many operational healthcare cases, discretization into a small action set is both clinically interpretable and statistically stable.

Offline evaluation often requires not only estimating value but also testing whether a candidate policy is better than a baseline. Define the improvement over baseline policy π_b as

$$\Delta(\pi, \pi_b)[33] = V(\pi) - V(\pi_b)$$

A robust evaluation can estimate Δ using paired estimators that share nuisance models to reduce variance. For example, a doubly robust estimator of Δ can be formed by taking the difference of two DR value estimates using the same $\widehat{\mu}$ and $\widehat{\pi}_0$. This reduces sensitivity to nuisance estimation noise and allows uncertainty quantification via influence functions.

Uncertainty quantification is essential for safety. For doubly robust estimators, asymptotic normality can be used under conditions, but healthcare datasets may violate these due to clustering and heavy tails from weighting. Bootstrap methods can be used, but naive bootstrap may fail under extreme weights [34]. A more robust approach uses influence-function-based variance estimates with weight trimming and cluster-robust corrections. Let the per-sample pseudo-outcome be

$$\phi_i(\pi) = \sum_{a \in \mathcal{A}} \widehat{\mu}(a, X_i)\, \pi(a \mid X_i) + \frac{\pi(A_i \mid X_i)}{\widehat{\pi}_0(A_i \mid X_i)} \Big[Y_i - \widehat{\mu}(A_i, X_i) \Big] [35]$$

Then $\widehat{V}_{\mathrm{DR}}(\pi) = \frac{1}{n}\sum_i \phi_i(\pi)$. A conservative variance estimate is

$$\widehat{\mathrm{Var}}\big(\widehat{V}_{\mathrm{DR}}(\pi)\big) = \frac{1}{n^2}\sum_{i=1}^{n}\Big(\phi_i(\pi) - \widehat{V}_{\mathrm{DR}}(\pi)\Big)^2$$

For clustered data with clusters $c = 1, \ldots, C$ and cluster index set $\mathcal{C}_c$, a cluster-robust analogue sums cluster-level residuals.

Offline policy testing must also handle constraints. In healthcare, an algorithm may be required to satisfy safety or equity constraints, such as limiting high-intensity outreach to patients with certain contraindications, or ensuring that a minimum fraction of high-risk patients receive nurse follow-up. Offline evaluation can incorporate constraints by estimating not only outcomes but also constraint-related quantities under the policy, such as expected resource usage [36]. Let G be a cost or resource random variable, and define

$$U(\pi) = \mathbb{E}\big[G(\pi)\big]$$

Then one can evaluate policies in the feasible set $\{\pi : U(\pi) \le u_{\max}\}$ using the same off-policy estimators with G in place of Y. This enables offline assessment of trade-offs between clinical utility and workload.

A final aspect of offline policy testing is counterfactual risk for harm. An algorithm might increase average utility while increasing risk in a vulnerable subgroup. Robust evaluation should therefore examine conditional value and tail risk, such as lower quantiles [37]. Estimating quantiles under off-policy weighting is difficult due to instability, but one can estimate conservative bounds using distributionally robust approaches. A practical alternative is to define harm indicators, such as whether a follow-up was delayed beyond a threshold, and evaluate their expected rate under the policy. While this still requires causal assumptions, it aligns with safety governance by focusing on measurable adverse events.

V. Robustness, Uncertainty, and Sensitivity Analysis

Robust evaluation requires explicit handling of three main fragilities: overlap limitations, unmeasured confounding, and distribution shift. Overlap limitations are visible in estimated propensities and weight distributions [38]. Unmeasured confounding is typically invisible but can be probed through sensitivity models. Distribution shift can be diagnosed by comparing covariate distributions and by evaluating the stability of nuisance models across time and sites.

Overlap can be assessed by examining $\widehat{\pi}_0(a \mid x)$ for actions used by the candidate policy. A practical diagnostic is the effective sample size under importance weights:

$$n_{\mathrm{eff}}(\pi) = \frac{\Big(\sum_{i=1}^{n} w(X_i, A_i)\Big)^2}{\sum_{i=1}^{n} w(X_i, A_i)^2}\,[39]$$

A small n_{eff} indicates that evaluation is dominated by a few high-weight samples, which is unsafe for decision-making. In healthcare deployments, it is often prudent to require n_{eff} above a threshold and to restrict policy changes in regions with poor overlap. This can be implemented by defining a restricted candidate policy that backs off to a baseline action when the estimated propensity under π_0 is below a minimum.

Weight stabilization can be improved by modeling the logging policy accurately. In healthcare, the logging policy can be influenced by site-specific protocols and clinician preferences. A hierarchical propensity model can capture site effects, while regularization prevents extreme probabilities. However, even a good propensity model cannot create overlap where none exists [40]. Therefore, robust evaluation should treat overlap constraints as design constraints, not merely as estimation nuisances.

Unmeasured confounding is a central concern. A common sensitivity approach introduces an unobserved variable U that affects both action and outcome. The goal is not to identify U but to quantify how large the effect of U would need to be to materially change conclusions. One approach models the deviation from ignorability through a selection bias function. For binary actions, consider the odds ratio of treatment assignment conditional on U: [41]

$$\mathrm{OR}(x) = \frac{\mathbb{P}(A=1 \mid X=x, U=1)\,\mathbb{P}(A=0 \mid X=x, U=0)}{\mathbb{P}(A=0 \mid X=x, U=1)\,\mathbb{P}(A=1 \mid X=x, U=0)}$$

A sensitivity parameter $\Gamma \ge 1$ can bound $\mathrm{OR}(x)$ by Γ for all x. One then derives bounds on the treatment effect or policy value consistent with that bound. In offline policy evaluation, this yields an interval of plausible policy values under different confounding strengths. For safety, the lower bound can be used as a conservative estimate, interpreting deployment as justified only if improvement persists under moderate confounding.

Another sensitivity approach uses outcome confounding functions. Suppose the true conditional mean differs from the observed regression by an additive bias term $b(a, x)$ due to unmeasured confounding: [42]

$$\mathbb{E}\big[Y(a) \mid X = x\big] = \mathbb{E}\big[Y \mid A = a, X = x\big] + b(a, x)$$

If one can bound $b(a, x)$ in magnitude, perhaps based on domain knowledge about unmeasured severity, then the policy value can be bounded. The policy value under bias becomes

$$V(\pi) = \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \big(\mu(a, X) + b(a, X)[43]\big)\,\pi(a \mid X)\Big]$$

In practice, one can explore a range of b forms, such as $b(a, x) = \delta_a$ constant shifts, or $b(a, x) = \delta_a s(x)$ where $s(x)$ is a severity score. The key is to report how conclusions change across plausible δ values rather than to select a single best guess.

Uncertainty quantification should reflect both sampling variability and nuisance estimation uncertainty [44]. Cross-fitting helps, but finite-sample inference can still be optimistic when weights are heavy-tailed. A robust approach uses trimming combined with sensitivity reporting. For example, one can report value estimates as a function of a weight cap c, producing a stability curve. If conclusions depend strongly on extreme weights, the evaluation is fragile. Similarly, one can report estimates across multiple nuisance model classes, such as generalized linear models, gradient boosting, and neural models, to assess model dependence [45]. The goal is not to average these models but to reveal when evaluation depends on a particular modeling choice.

Distribution shift is a major concern in healthcare personalization because deployment changes the action distribution and potentially the covariate distribution through feedback. Even without feedback, the patient population can drift. Offline evaluation typically assumes that the conditional outcome model $\mathbb{E}[Y \mid X, A]$ is stable between the logged environment and the deployment environment. If this invariance fails, the offline estimate may not transport. One robustness strategy is to evaluate policies on temporally separated validation sets, for example training nuisance models on earlier months and evaluating on later months, to assess stability. Site-based validation, where models are trained on some facilities and evaluated on others, can test generalization across operational contexts [46]. While these tests are not definitive for causal transportability, they provide practical evidence about the stability of predictive relationships that underpin causal estimators.

A formal approach to covariate shift introduces a target distribution $p^\star(x)$ and defines the target policy value

$$V^\star(\pi) = \mathbb{E}_{X \sim p^\star}\Big[\sum_{a \in \mathcal{A}} \mathbb{E}\big[Y(a) \mid X\big]\, \pi(a \mid X)\Big]$$

If one can estimate the density ratio $r(x) = p^\star(x)/p(x)$, then one can reweight samples to approximate the target [47]. Combining this with off-policy weights yields

$$\widehat{V}^\star_{\mathrm{IS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)\, w(X_i, A_i)\, Y_i$$

This estimator can be extremely unstable because it multiplies two ratios. Therefore, robust evaluation in practice often restricts to moderate shifts or uses conservative uncertainty [48]. Another approach is to define the target population through explicit inclusion criteria and to evaluate only on that subset, avoiding extreme density ratios.

Missing outcomes are pervasive. Suppose M is an indicator that Y is observed. If missingness depends on covariates and action, and if it is missing at random given (X, A), then one can use inverse probability of observation weighting. Let $\rho(x, a) = \mathbb{P}(M = 1 \mid X = x, A = a)$. Then a doubly robust estimator for observed outcomes can incorporate both off-policy weights and observation weights:

$$\widehat{V}_{\mathrm{DR,MO}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a \in \mathcal{A}} \widehat{\mu}(a, X_i)\, \pi(a \mid X_i)$$

[49] $$+ \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\widehat{\pi}_0(A_i \mid X_i)}\, \frac{M_i}{\widehat{\rho}(X_i, A_i)}\Big[Y_i - \widehat{\mu}(A_i, X_i)\Big]$$

This formulation highlights that healthcare evaluation can require multiple layers of weighting, which amplifies variance. Therefore, robust evaluation must treat missingness modeling and overlap diagnostics as central rather than peripheral.

Finally, robustness includes fairness and subgroup stability. Even if the average value improves, a policy might worsen outcomes for specific protected or clinically vulnerable groups [50]. One can define subgroup-specific value

$$V_g(\pi) = \mathbb{E}\big[Y(\pi) \mid G = g\big]$$

where G is a group indicator. Estimating $V_g(\pi)$ off-policy can be done by applying the same estimators within each group or by including group interactions in nuisance models. However, subgroup evaluation often has smaller effective sample sizes and worse overlap, requiring additional conservatism. A robust governance approach uses subgroup lower bounds or constraints that require non-inferiority within groups, recognizing that statistical uncertainty may be substantial [51].

VI. Operational Considerations in Healthcare Deployment

Robust evaluation is not purely a statistical exercise. Healthcare personalization systems operate within privacy regulations, clinical governance, and operational constraints. Offline policy testing must align with data governance, auditability, and safety monitoring. This section translates the causal and off-policy methods into an implementable workflow suitable for healthcare organizations while recognizing that details vary across jurisdictions and care settings.

Data governance affects feature availability and logging quality. Privacy regulations and organizational policies may limit the use of sensitive attributes, but robust evaluation often benefits from including confounders that reduce bias [52]. A pragmatic approach is to separate the modeling pipeline into a confounding-adjustment layer and a decision layer, where sensitive features may be used for adjustment and auditing but not for direct targeting, depending on policy. This separation requires careful controls to prevent leakage and to ensure that the deployed policy complies with governance rules. Regardless of whether sensitive attributes are used in the policy, subgroup evaluation often requires them for fairness auditing and safety checks.

Logging is critical. Offline evaluation relies on accurate records of which action was taken, when, and under what

context [53]. In many healthcare settings, actions occur across heterogeneous systems such as electronic health records, scheduling platforms, and communication vendors. A robust evaluation pipeline should define a canonical event schema with immutable timestamps and consistent identifiers. Without this, positivity diagnostics and propensity estimation can be distorted by inconsistent action coding. Furthermore, the logging policy may change over time due to operational updates, requiring versioning of π_0 and time-stratified evaluation.

Clinical constraints define the action set. Some actions are clinically contraindicated for certain patients [54]. These contraindications should be encoded explicitly as feasibility constraints, not learned implicitly from data. If contraindications are not encoded, an algorithm may propose unsafe actions that cannot be evaluated offline because they never occurred, and might also create spurious overlap issues. Feasibility constraints also help interpret positivity: a zero propensity due to clinical constraint is acceptable because the candidate policy should not choose that action. A robust evaluation should therefore define $\mathcal{A}(x)$ as the feasible action set given x, and restrict policies to $\pi(a \mid x) = 0$ for $a \notin \mathcal{A}(x)$.

Capacity constraints and resource allocation matter. For example, nurse calls are limited. A policy that increases calls to high-risk patients might improve outcomes but exceed staffing [55]. Offline evaluation can estimate expected resource usage as described earlier, but true operational feasibility may depend on temporal clustering and peak loads. A policy that concentrates calls at specific times may overload staff even if average volume is acceptable. Therefore, robust evaluation should incorporate temporal features and simulate operational queues when possible, treating them as part of the environment. When detailed simulation is not available, conservative constraints that limit the fraction of patients assigned to high-cost actions can reduce risk.

Safety monitoring requires metrics beyond average outcomes. Healthcare deployments typically require monitoring of adverse events, process measures, and equity indicators [56]. Offline evaluation can pre-specify these metrics and estimate their expected values under candidate policies. However, post-deployment monitoring remains necessary because offline estimates can fail under shift. A robust evaluation report should therefore include trigger conditions, such as detecting a drop of more than $x\%$ in appointment completion for a subgroup, or an increase in emergency visits beyond a threshold, with clearly defined response actions such as rolling back to a baseline policy. While the monitoring plan is operational rather than mathematical, it is part of robust evaluation because it defines how residual uncertainty is managed.

Model risk management benefits from reproducibility and audit trails. Because causal evaluation depends on modeling choices and assumptions, the pipeline should record data versions, feature construction logic, propensity and outcome model configurations, weight trimming thresholds, and sensitivity parameter ranges [57]. In regulated environments, these artifacts support internal review and external audits. They also support iterative improvement as more data becomes available.

A common operational pattern is to compare several candidate policies, including a baseline, a conservative incremental change, and a more aggressive personalization. Offline evaluation can rank these policies under estimated value and uncertainty. However, robust decision-making should avoid selecting a policy solely on point estimates. In healthcare, it is often preferable to select a policy whose lower confidence bound exceeds the baseline or whose sensitivity-adjusted lower bound remains favorable under moderate confounding [58]. This can be viewed as a robust optimization criterion where the deployment choice maximizes a conservative value estimate rather than the mean estimate.

Evaluation should also consider patient experience and burden. Actions such as repeated reminders may improve appointment completion but reduce satisfaction or increase opt-out rates. If such outcomes are not included in Y, the policy evaluation may be incomplete. A robust approach defines Y as a composite utility or conducts multi-outcome evaluation, ensuring that improvements in one metric do not hide harms in another. When outcomes are hard to measure, proxy measures such as opt-out rates, complaint tickets, or call abandonment can be incorporated [59].

Finally, offline evaluation should respect the difference between internal validity and external validity. Even if identification assumptions are plausible in a logged dataset, they may not hold in other sites or populations. Therefore, robust evaluation should stratify by site, payer, and care setting when possible, and should avoid extrapolation beyond overlap. If an organization plans to deploy to a new clinic with different workflows, offline evaluation from existing sites should be treated as informative but not definitive. In such cases, a cautious deployment strategy might involve limited rollout with enhanced monitoring, even when full randomized trials are not feasible [60].

VII. Mathematical Framework for Robust Estimation Under Practical Constraints

This section consolidates the key estimators into a unified mathematical view, emphasizing robustness to nuisance estimation, overlap limitations, and partial identification. The intent is to clarify how different components interact in

healthcare settings where data is high-dimensional, actions are constrained, and outcomes are noisy.

Let P denote the joint distribution of (X, A, Y) under the logging policy. Let π be a candidate policy. Under identification assumptions, the policy value functional can be written as

$$V(\pi)[61] = \mathbb{E}_P\Big[\sum_{a\in\mathcal{A}} \mu(a, X)\,\pi(a \mid X)\Big]$$

$$\mu(a, x) = \mathbb{E}_P\big[Y \mid A = a, X = x\big]$$

Define the propensity $\pi_0(a \mid x) = \mathbb{P}_P(A = a \mid X = x)$. Consider the efficient influence function in the nonparametric model for $V(\pi)$, which motivates doubly robust estimation:

$$\varphi(Z; \pi, \mu, \pi_0)[62] = \sum_{a\in\mathcal{A}} \mu(a, X)\,\pi(a \mid X) + \frac{\pi(A \mid X)}{\pi_0(A \mid X)}\Big[Y - \mu(A, X)\Big] - V(\pi)$$

where $Z = (X, A, Y)$ [63]. A plug-in estimator replaces $(\mu, \pi_0, V(\pi))$ with estimates and uses the sample mean, yielding the DR estimator presented earlier. Cross-fitting mitigates overfitting bias by ensuring that $\widehat{\mu}$ and $\widehat{\pi}_0$ are estimated on data independent of the evaluation fold.

In healthcare, overlap issues motivate restricting policies to regions where π_0 is not too small. Define an overlap indicator

$$\Omega(X) = \mathbb{I}\Big\{\min_{a:\,\pi(a|X)>0} \pi_0(a \mid X) \geq \epsilon[64]\Big\}$$

for some $\epsilon > 0$. One can define a restricted value

$$V_\epsilon(\pi) = \mathbb{E}\big[Y(\pi)\,\Omega(X)\big] \,/\mathbb{E}\big[\Omega(X)\big]$$

which evaluates the policy only on contexts with sufficient overlap. Operationally, this corresponds to deploying the policy only where the system has evidence, and backing off elsewhere. Estimating $V_\epsilon(\pi)$ is more stable, though it changes the target estimand. A robust evaluation report should state the coverage $\mathbb{E}[\Omega(X)]$ to quantify what fraction of decisions are supported by overlap.

Weight trimming can be formalized as a constrained estimator [65]. Let w_i be the importance weight. Define trimmed weights

$$\widetilde{w}_i = \min\big(w_i, c\big)$$

and consider a trimmed DR estimator where the residual term uses $\widetilde{w}_i$. This can be interpreted as evaluating the policy under a contamination model where extreme ratios are treated as unreliable. A robustness analysis can report how $\widehat{V}$ changes across c. If the estimate is stable across a broad range of c, conclusions are less sensitive to overlap fragility.

Healthcare outcomes often exhibit heavy tails, for example in cost outcomes or utilization counts [66]. In such cases, robust M-estimation can be used to reduce sensitivity to outliers. One can replace the residual term in DR with a bounded influence function ψ, such as Huber's function:

$$\widehat{V}_{\mathrm{DR},\psi}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \widehat{\mu}(a, X_i)\,\pi(a \mid X_i) + \frac{1}{n}\sum_{i=1}^{n} w(X_i, A_i)\,[67]\psi\Big(Y_i - \widehat{\mu}(A_i, X_i)\Big)$$

This introduces bias if ψ is nonlinear, but can improve stability when outcome noise is extreme. In safety-critical decision-making, stability can be prioritized, with bias assessed via sensitivity analyses and validation against held-out time periods.

Partial identification arises when ignorability is not fully credible. In that case, instead of a point value $V(\pi)$, one can aim to bound it [68]. Consider a binary action for simplicity and define a sensitivity parameter $\eta(x)$ that bounds the difference between the true propensity and the observed propensity due to unmeasured confounding. One stylized model is

$$\mathbb{P}(A = 1 \mid X = x, Y(1), Y(0)) = \mathrm{logit}^{-1}\Big(\mathrm{logit}\big(\pi_0(1 \mid x)\big) + \eta(x)\,\Delta_Y\Big)$$

$$\Delta_Y[69] = Y(1) - Y(0)$$

When $\eta(x) = 0$, ignorability holds. As $\eta(x)$ increases, assignment depends more strongly on the individual treatment effect. This model implies that patients more likely to benefit may be more likely to receive the action, or vice versa. Although Δ_Y is unobserved, one can derive bounds on policy value by optimizing over distributions consistent with observed data and η constraints. In practice, exact optimization can be complex, but approximate bounds can be computed by exploring plausible ranges of η and assessing whether conclusions about policy ranking change [70]. The operational message is that robust evaluation should articulate how strong unmeasured confounding would need to be to reverse a deployment decision, rather than treating ignorability as a binary yes or no.

For sequential decisions, robust evaluation can use marginal importance sampling with per-time-step weights. Let $\pi_{0,t}(a_t \mid h_t)$ be the logging policy at time t, and π_t the candidate. The trajectory weight is

$$W = \prod_{t=1}^{T} \frac{\pi_t(A_t \mid H_t)}{\pi_{0,t}(A_t \mid H_t)}$$

Trajectory weights can explode with T, making naive importance sampling unusable. Practical healthcare implementations often use truncated horizons, per-decision evaluation, or doubly robust sequential estimators based on recursive value models. Even then, robustness requires careful overlap assessment at each time [71]. This motivates designing

personalization systems with evaluation in mind, including logging stochasticity in safe regions to improve support and enable learning.

Taken together, the mathematical structure suggests that robust offline evaluation is less about selecting a single estimator and more about managing the fragilities exposed by the estimators. Overlap determines whether weights are stable. Ignorability determines whether the estimand is meaningful. Model estimation determines whether DR properties hold in finite samples. Robust evaluation therefore requires a multi-pronged approach: estimator ensembles, trimming and restriction for overlap, sensitivity bounds for confounding, and temporal or site validation for shift [72].

VIII. Conclusion

Robust evaluation of customer personalization algorithms in healthcare requires integrating causal inference with offline policy testing in a way that respects clinical constraints, data limitations, and safety obligations. The key target is not a predictive metric but the counterfactual policy value, defined as the expected outcome that would arise if a candidate personalization policy were deployed. Identification of this value from logged observational data relies on consistency, conditional exchangeability, and positivity, assumptions that are often strained in healthcare due to latent severity, clinician judgment, operational constraints, and incomplete measurement. Offline policy estimators based on importance sampling, direct outcome modeling, and doubly robust constructions provide practical tools for estimating policy value, but their reliability depends critically on overlap and nuisance model quality.

A robust framework therefore treats diagnostics and sensitivity analysis as essential components rather than optional add-ons. Overlap should be quantified through propensity diagnostics, weight distributions, and effective sample size, with policy restriction or backoff strategies when support is inadequate [73]. Unmeasured confounding should be addressed through sensitivity models and bounds that clarify how conclusions depend on unverifiable assumptions. Distribution shift should be probed through temporal and site-based validation, and missingness mechanisms should be modeled explicitly when outcomes are partially observed. Uncertainty quantification should be conservative, reflecting heavy-tailed weights, clustering, and finite-sample variability, and decision-making should emphasize lower bounds and stability across modeling choices rather than maximizing point estimates.

Operationally, robust evaluation must be embedded in governance structures that define feasibility constraints, safety metrics, auditing practices, and monitoring triggers. Healthcare personalization is a socio-technical system: evaluation results are only as actionable as the logging infrastructure, the clarity of outcome definitions, and the organization's capacity to monitor and respond to post-deployment deviations. Within these constraints, causal offline policy testing can support comparative assessment of candidate algorithms without requiring immediate live experimentation, while making explicit the assumptions and uncertainties that determine how much confidence the organization should place in offline evidence [74].

REFERENCES

[1] N. Kumari and P. K. S. Kumar, "Customer satisfaction of pharmacy services of tertiary care hospital: A review," *International Journal of Health Sciences and Pharmacy*, pp. 128–148, 6 2023.

[2] S. J. Moschuris and M. N. Kondylis, "Outsourcing in private healthcare organisations: a greek perspective.," *Journal of health organization and management*, vol. 21, pp. 220–223, 6 2007.

[3] S. A. Kurtuluş and E. Cengiz, "Customer experience in healthcare: Literature review," *Istanbul Business Research*, vol. 0, pp. 0–0, 12 2019.

[4] M. R. Karim, N. Nordin, M. S. Hassan, M. A. Islam, and M. F. Yusof, "Erp in bangladesh healthcare industry: Current scenario, challenges, and future directions.," *Work (Reading, Mass.)*, vol. 81, pp. 2010–2020, 1 2025.

[5] D. N. Jumapili and S. M. A. Muathe, "Quality management practices and performance: The perspective of public healthcare institutions in kenya," *European Scientific Journal, ESJ*, vol. 21, pp. 83–83, 2 2025.

[6] S. H. Kukkuhalli, "Enabling customer 360 view and customer touchpoint tracking across digital and non-digital channels," *Journal of Marketing & Supply Chain Management*, vol. 1, no. 3, 2022.

[7] V. R. Pereira, M. Kreye, and M. M. de Carvalho, "Customer-pulled and provider-pushed pathways for product-service system: The contingent effect of the business ecosystems," *Journal of Manufacturing Technology Management*, vol. 30, pp. 729–747, 6 2019.

[8] T. Kortana, "Key success for registered nurses to be entrepreneur in senior healthcare business in thailand," *Chinese Business Review*, vol. 16, 7 2017.

[9] N.-H. Duong and X. M. N. Chau, "The impact of social media communication and customer experience in healthcare services: A case study of tam anh general hospital," *Journal of Development and Integration*, pp. 26–35, 10 2025.

[10] S. H. K. Hosseini and L. Behboudi, "Brand trust and image: effects on customer satisfaction," *International journal of health care quality assurance*, vol. 30, pp. 580–590, 8 2017.

[11] "Influence of ai on shopping experience of generation z customers," *Recent trends in Management and Com-*

merce, vol. 5, pp. 8–13, 7 2024.

[12] D. D. V. Fleet and T. O. Peterson, "Improving healthcare practice behaviors: An exploratory study identifying effective and ineffective behaviors in healthcare," *International journal of health care quality assurance*, vol. 29, pp. 141–161, 3 2016.

[13] M. S. K, "A study on inventory management strategies for medical equipment suppliers," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 9, pp. 1–9, 4 2025.

[14] R. R. S. D. Wulandari, A. Khatibi, S. M. F. Azam, J. Tham, and C. Widiyaningsih, "The effect of marketing mix on outpatients loyalty in hajj hospital jakarta, indonesia," *Saudi Journal of Business and Management Studies*, vol. 8, pp. 267–277, 11 2023.

[15] I. Gremyr and H. Raharjo, "Quality function deployment in healthcare: a literature review and case study," *International journal of health care quality assurance*, vol. 26, pp. 135–146, 2 2013.

[16] O. Iparraguirre-Villanueva, L. Obregon-Palomino, W. Pujay-Iglesias, F. Sierra-Liñan, and M. Cabanillas-Carbonell, "Productivity of incident management with conversational bots-a review," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, pp. 1543–1543, 12 2023.

[17] A. Singh, A. Jha, and S. Purbey, "Drivers of home healthcare industry: Contrasting views of users and providers in a developing country," *Journal of Pharmaceutical Research International*, pp. 72–80, 7 2021.

[18] S. Mallireddy, "What are best tools for digital transformation in today world helping people from health and banking sectors," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 8, pp. 1–7, 10 2024.

[19] G. Hiermann, M. Prandtstetter, A. Rendl, J. Puchinger, and G. R. Raidl, "Metaheuristics for solving a multimodal home-healthcare scheduling problem," *Central European Journal of Operations Research*, vol. 23, pp. 89–113, 5 2013.

[20] D. F. Yulianto, T. Pratiwi, F. A. Irsan, and F. Mustikasari, "Consumer behavior switching from human agents to chatbots in the health service industry," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 14, pp. 355–355, 8 2025.

[21] P. Pawar and K. Rodrigues, "Walk more, pay less: A study of customer awareness, engagement, and comparative analysis of the wellness payback model in retail health insurance," *International Journal For Multidisciplinary Research*, vol. 7, 12 2025.

[22] J. R. Machireddy, "A two-stage ai-based framework for determining insurance broker commissions in the healthcare industry," *Transactions on Artificial Intelligence, Machine Learning, and Cognitive Systems*, vol. 7, no. 6, pp. 1–21, 2022.

[23] M. Delima, S. Hanoum, and M. S. Salahudin, "Customer satisfaction analysis in the healthcare industry using kano method and importance performance analysis," *Jurnal Teknologi dan Manajemen*, vol. 6, pp. 81–90, 7 2025.

[24] Y. J. Chamariya, ""revolutionizing medical insurance with ai/ml integration"," *Nanotechnology Perceptions*, vol. 17, pp. 289–298, 10 2021.

[25] R. Brander, M. Paterson, and Y. E. Chan, "Fostering change in organizational culture using a critical ethnographic approach," *The Qualitative Report*, vol. 17, pp. 1–27, 1 2015.

[26] F. Ponsignon, A. Smart, and L. Phillips, "A customer journey perspective on service delivery system design: insights from healthcare," *International Journal of Quality & Reliability Management*, vol. 35, pp. 2328–2347, 11 2018.

[27] S. I. Hoque, A. M. Karim, R. Hossen, and D. Arjumand, "Evaluation of patients' satisfaction in telemedicine service quality: A case study on maizbhanderi foundation, fatikchari, bangladesh," *American Economic & Social Review*, vol. 8, pp. 1–10, 8 2021.

[28] N. A. M. Rosli, A. Suleiman, V. Arumugam, and M. R. Ismail, "Conceptual development on factors influencing customers satisfaction in private rehabilitation hospitals in klang valley," *International Journal For Multidisciplinary Research*, vol. 5, 11 2023.

[29] R. Kumar and P. Jaiswal, "Impact of digital marketing on online drugs and healthcare products sales in bihar," *International Journal For Multidisciplinary Research*, vol. 5, 4 2023.

[30] Y. Yang, "Does economic growth induce smoking?—evidence from china," *Empirical Economics*, vol. 63, no. 2, pp. 821–845, 2022.

[31] M. M. Willie, "Challenges and considerations in assessing healthcare service quality: A comprehensive analysis," *Golden Ratio of Marketing and Applied Psychology of Business*, vol. 4, pp. 152–160, 6 2023.

[32] A. Moreira, J. Duarte, and M. F. Santos, "Case study of multichannel interaction in healthcare services," *Information*, vol. 14, pp. 37–37, 1 2023.

[33] L. Eberle, G. S. Milan, and D. D. Toni, "Dimensiones del riesgo percibido en el proceso decisorio de elecciÓn de un plan de salud privado: Un estudio de caso," *Revista Facultad de Ciencias Económicas*, vol. 22, pp. 49–61, 6 2014.

[34] A. Y. Aleryani, "Assessing the quality of health content on arabic digital platforms: Insights from healthcare professionals and the public," *European Modern Studies Journal*, vol. 7, pp. 71–86, 11 2023.

[35] M. Watson, "Panel discussion on the application of mcda tools.," *Cost effectiveness and resource allocation : C/E*, vol. 16, pp. 40–40, 11 2018.

[36] N. F. Habidin, "The development of lean healthcare management system (lhms) for healthcare industry,"

Asian Journal of Pharmaceutical and Clinical Research, vol. 10, pp. 97–102, 2 2017.

[37] L. F. Restrepo and E. C. Paredes, "Theoretical of macroeconomic conditions for tobacco control: Affordability constraints vs. psychological distress mechanisms," *Quarterly Journal of Emerging Scientific Trends and Technologies*, vol. 14, no. 4, pp. 1–10, 2024.

[38] T. W. Mgaya and R. Gwahula, "The effect of service quality on customer satisfaction at st. joseph referral hospital-peramiho, tanzania," *European Journal of Theoretical and Applied Sciences*, vol. 2, pp. 223–233, 3 2024.

[39] S. H. Kukkuhalli, "Improving digital sales through reducing friction points in the customer digital journey using data engineering and machine learning," *International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences*, vol. 10, no. 3, 2022.

[40] I. V. Pestun and Z. M. Mnushko, " '," *Socìalna farmacìâ v ohoronì zdorov'â*, vol. 2, pp. 57–63, 3 2016.

[41] P. Anabila, D. K. Kumi, and J. Anome, "Patients' perceptions of healthcare quality in ghana.," *International journal of health care quality assurance*, vol. 32, pp. 176–190, 2 2019.

[42] A. Y. Pratama, D. Andiyana, M. M. Akbar, N. P. T. Fenriyanti, and Y. Prasetya, "Marketing mix strategy (7ps) in building user loyalty in specialized hospitals and general hospitals: A systematic review of various service sectors with implications for healthcare services," *Formosa Journal of Multidisciplinary Research*, vol. 4, pp. 5875–5892, 12 2025.

[43] M. K. Poku, N. A. Behkami, and D. W. Bates, "Patient relationship management: What the u.s. healthcare system can learn from other industries.," *Journal of general internal medicine*, vol. 32, pp. 101–104, 8 2016.

[44] M. Usman, M. Mujahid, F. Rustam, E. Flores, J. L. V. Mazón, I. de la Torre Díez, and I. Ashraf, "Analyzing patients satisfaction level for medical services using twitter data.," *PeerJ. Computer science*, vol. 10, pp. e1697–e1697, 1 2024.

[45] J. C. Lee, "The perils of artificial intelligence in healthcare: Disease diagnosis and treatment," *Journal of Computational Biology and Bioinformatics Research*, vol. 9, pp. 1–6, 4 2019.

[46] C. P. Ratna, R. T. Juwaheer, and S. Pudaruth, "Assessing the impact of technology adoption on human touch aspects in healthcare settings in mauritius," *Studies in Business and Economics*, vol. 13, pp. 164–178, 8 2018.

[47] D. Min, "Analysis of customer arrival process in an appointment-based service system," *Journal of the Korean Operations Research and Management Science Society*, vol. 37, pp. 31–43, 6 2012.

[48] M. F. Elsayed, "Conceptualizing the nexus between aggregate economic shocks, socioeconomic status, and persistent smoking addiction," *Annual Compendium of Multidisciplinary Scientific Insights and Technological Frontiers*, vol. 14, no. 3, pp. 1–14, 2024.

[49] A. M. Jain, "Ai-powered business intelligence dashboards : A cross-sector analysis of transformative impact and future directions," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 524–536, 7 2023.

[50] M. Ghabban, "The impact of marketing mix (7ps) on customer satisfaction in the healthcare sector," *International Journal of Healthcare Information Systems and Informatics*, vol. 20, pp. 1–30, 2 2025.

[51] J. Hermawan, A. Rianawati, C. R. S. Prakoeswa, and L. I. Wijaya, "Leveraging it use in healthcare records management: impact on customer satisfaction and financial performance through knowledge transfer and absorptive capacity," *Business: Theory and Practice*, vol. 26, pp. 323–334, 9 2025.

[52] T. Jayasekara and L. Weerasinghe, "Comparison of instrumental variable approaches in addressing endogeneity within linear panel data models," *International Journal of Advanced Scientific Computation, Modeling, and Simulation*, vol. 14, no. 7, pp. 1–21, 2024.

[53] S. R. Velamuri, P. Anant, and V. Kumar, "Doing well to do good: Business model innovation for social healthcare," *Business Models and Modelling*, vol. 33, pp. 279–308, 11 2015.

[54] D. A. Adeyemi, N. C. Amedior, B.-P. Longe, F. Akintunwa, S. Omotosho, and O. B. Longe, "Ethical implications of artificial intelligence in the global healthcare sector: Trends, perspectives and future developments," *Advances in Multidisciplinary & Scientific Research Journal Publications*, vol. 9, pp. 41–53, 9 2023.

[55] J. R. Machireddy, "Automation in healthcare claims processing: Enhancing efficiency and accuracy," *International Journal of Science and Research Archive*, vol. 09, no. 01, pp. 825–834, 2023.

[56] H. S. S. Albalwei and F. M. Albalawi, "Dimensions of quality in health care facilities: A simple review article," *Asian Journal of Medicine and Health*, pp. 11–16, 5 2022.

[57] A. Dobele and A. Lindgreen, "Exploring the nature of value in the word-of-mouth referral equation for health care," *Journal of Marketing Management*, vol. 27, pp. 269–290, 2 2011.

[58] W. Xu, "Optimizing product design, marketing and customer relationship management in healthcare enterprises through data analytics - a case study based on a healthcare enterprise," *Advances in Economics, Management and Political Sciences*, vol. 149, pp. 180–189, 1 2025.

[59] C. P. Price, A. S. John, R. H. Christenson, V. Scharnhorst, M. Oellerich, P. M. Jones, and H. A. Morris, "Leveraging the real value of laboratory medicine with the value proposition," *Clinica chimica acta; interna-*

tional journal of clinical chemistry, vol. 462, pp. 183–186, 9 2016.

[60] A. Saputra and H. Wijaya, "The influence of trade liberalization policies on pharmaceutical and medical technology pricing and public health accessibility in global supply chains," *Quarterly Journal of Emerging Scientific Trends and Technologies*, vol. 15, no. 4, pp. 1–13, 2025.

[61] V. Rodriguez-Trujillo and Y. F. Ceballos, "Análisis de la variación de los usuarios en entidades de medicina prepagada en colombia – caso de estudio," *AiBi Revista de Investigación, Administración e Ingeniería*, vol. 8, pp. 124–130, 1 2020.

[62] S. Bellary, P. K. Bala, and S. Chakraborty, "Utilizing online reviews for analyzing digital healthcare consultation services: Examining perspectives of both healthcare customers and healthcare professionals.," *International journal of medical informatics*, vol. 191, pp. 105587–105587, 8 2024.

[63] J. R. Machireddy, "Data science and business analytics approaches to financial wellbeing: Modeling consumer habits and identifying at-risk individuals in financial services," *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, vol. 7, no. 12, pp. 1–18, 2023.

[64] M. Ileana, P. Petrov, and V. Milev, "Secure integration of sensor networks and distributed web systems for electronic health records and custom crm.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 5102–5102, 8 2025.

[65] E. Ardava, N. Gustiņa, R. Šukele, and O. Onževs, "Pharmaceutical care and effective communication with customers," *SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference*, vol. 4, pp. 294–306, 5 2021.

[66] A. M. Díaz-Martín, A. Schmitz, and M. J. Y. Guillén, "Are health e-mavens the new patient influencers?," *Frontiers in psychology*, vol. 11, pp. 779–, 4 2020.

[67] T. Win and E. E. Phyo, "Making the appointment at opd by using clustering algorithm," *International Journal of Science and Engineering Applications*, vol. 8, pp. 231–235, 7 2019.

[68] N. Yirdew, K. A. Mnoj, and L. Bhaskari, "Privacy enhancement in cloud environment through trusted third party approach," *International Journal of Advanced Research in Computer Science*, vol. 8, pp. 273–277, 8 2017.

[69] S. K. Rosahl and M. Samii, ""nihil nocere" and beyond: The issue of leadership in healthcare," *HealthcarePapers*, vol. 4, pp. 78–82, 7 2003.

[70] V. Scafarto, T. Dalwai, F. Ricci, and G. della Corte, "Digitalization and firm financial performance in healthcare: The mediating role of intellectual capital efficiency," *Sustainability*, vol. 15, pp. 4031–4031, 2 2023.

[71] S. H. Kukkuhalli, "Increasing digital sales revenue through 1:1 hyper-personalization with the use of machine learning for b2c enterprises," *Journal of Artificial Intelligence, Machine Learning and Data Science*, vol. 1, no. 1, 2023.

[72] L. Braun, E. Tiralongo, J. M. Wilkinson, S. Poole, O. Spitzer, M. Bailey, and M. J. Dooley, "Adverse reactions to complementary medicines: the australian pharmacy experience.," *The International journal of pharmacy practice*, vol. 18, pp. 242–244, 7 2010.

[73] K. S. Hong and D. Lee, "Impact of operational innovations on customer loyalty in the healthcare sector," *Service Business*, vol. 12, pp. 575–600, 11 2017.

[74] A. Singh and A. Prasher, "Measuring healthcare service quality from patients' perspective: using fuzzy ahp application," *Total Quality Management & Business Excellence*, vol. 30, pp. 284–300, 3 2017.