stats

MDPI

Article

Sharp Estimand-Specific Precision Envelopes for Blockwise Missing Data Under Missing at Random

Satoshi Nakamura

Department of Advanced Diagnostic Imaging, Mie University Graduate School of Medicine, 2-174 Edobashi, Tsu 514-8507, Mie, Japan; s-nakamura@med.mie-u.ac.jp

Abstract

An overall missingness rate does not determine how much precision survives: at 25% missing, an exact calculation raises the standard error by 1.2% under one admissible mechanism and 21.4% under another. For variable or all-or-none block missingness at random, we characterize the precision attainable across mechanisms sharing an observation rate and a lower bound on observation probabilities. The envelope is sharp: it contains every attainable efficiency bound. The favorable endpoint is explicit. In atomless spaces the adverse endpoint integrates the quantile function of the residual influence variance, so it is evaluated, not optimized; in mixed atomic and nonatomic spaces it has an exact attained variational characterization. Crossings with a precision requirement divide missingness into guaranteed, allocation-sensitive, and inadmissible regions. The construction also handles absolute variance and interval width requirements and returns the minimum-cost recovery as a priority ordering. For a partially observed continuous outcome, cross-fitted one-step estimators and a fail-closed bootstrap give asymptotic simultaneous inference on a prespecified grid under stated regularity conditions, and inversion yields an ordered confidence set for both thresholds. Without an overlap restriction, no positive observation rate guarantees nontrivial retention. Public clinical illustrations are reported, with Monte Carlo joint coverage below nominal in the smallest samples.

Keywords: efficient influence function; information loss; missing at random; positivity; relative efficiency; sharp bounds

1. Introduction

Whether a study has too much missing data is commonly settled by a rule of thumb about the proportion missing. Consider, for example, a cohort prepared for analysis in which the missing proportions differ from variable to variable, and one patient in five lacks the measurement the planned analysis takes as its outcome. The team applies the familiar standard that missingness of this order is tolerable, and the reviewer of the report will apply a standard of the same kind. In such a situation, the investigator might ask: is there a proportion below which missingness is safe?

No proportion is safe on its own. The proportion missing predicts imputation performance poorly, response rates diagnose nonresponse bias poorly, and analyses that weight by the inverse of an observation probability grow unstable when some of those probabilities approach zero, because which units went unobserved matters [1–3]. Published percentage guidance accordingly depends on the outcome and on the method used [4]. Moreover, one rate does not mean the same thing for two estimands. What unobserved unit costs depend on is how much it would have influenced the target, so a rate can be tolerable

Check for updates

Academic Editor: Wei Zhu

Received: 15 September 2026
Revised: 30 September 2026
Accepted: 1 October 2026
Published: 3 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

for a descriptive summary and disqualifying for the contrast that the study was designed to make on the same rows. Multiple imputation and likelihood methods support valid inference under appropriate assumptions [5–7], but they analyze the pattern that occurred and say nothing about how much information that pattern left.

That last question is easily run together with two others. The first asks whether the remaining data still identify the target, and the second asks which estimator or imputation procedure to run. This article asks neither. It asks how much of the complete-data information can survive at a fixed observation rate, for one variable or one all-or-none block, assuming that data missing at random (MAR) identify the estimand and that an efficient estimator exists. The article does not test MAR and does not certify any imputation algorithm.

The answer takes the form of a decision about a given rate, and it rests on the following four inputs: the estimand, a precision requirement expressed either as a fraction of the complete-data information or as a variance or confidence interval width, a lower bound on the probability that any unit is observed, and the observation rate itself. From these the article computes the precision envelope, that is, the range that the rate permits, running from the most favorable admissible mechanism to the least favorable one. Comparing the envelope with the requirement assigns the rate to one of three phases. In the guaranteed phase the requirement holds whatever the mechanism; in the allocation-sensitive phase it holds under some admissible mechanisms and fails under others; and in the inadmissible phase no mechanism at that rate can meet it.

When the requirement fails and the planning inputs stay fixed, the same calculation ranks individual units by the precision that observing one more of them would buy per unit of cost, so the reply names which observations to add instead of naming a lower overall percentage. Applied to data, these quantities are estimated by cross-fitting, and the phase assignment carries a simultaneous confidence statement. One implementation carries individual records from a data file through to that decision.

The efficiency bound used throughout is the standard locally nonparametric bound under MAR [8,9]. Choosing a favorable allocation descends from Neyman allocation and two-phase design, and holding the allocation above a probability floor gives a box-constrained optimization problem [10–14]; active acquisition rests on related principles [15,16]. Those methods select one favorable rule under a modeled mechanism. The envelope instead ranges over every admissible mechanism, so it needs no model for the observation probabilities and no smoothness or atomlessness in the covariate space, and its endpoints depend on the complete-data law only through a single function that can be estimated. The new object is the uncertainty set that these components combine to produce once the observation rate and a common probability floor are fixed. Closest to the present objective is the two-phase sampling rule of Wang et al. [17]. For a scalar target they choose a design that minimizes a semiparametric efficiency bound, and for a vector target a maximin rule that hedges across its components, so their design resolves uncertainty about which estimation task it must serve. Here, the estimand is fixed and the uncertainty concerns the mechanism: the envelope selects no design and returns the whole range of bounds the admissible class produces at a given observation rate, in mixed atomic and nonatomic spaces alike. A phase certification rests on that range.

The estimation side likewise draws on existing work. The two variance components that determine retention are analysis of variance measures of variable importance, and their one-step estimation follows Williamson et al. [18], while the translation into an equivalent complete-data sample size extends familiar effective sample size calculations [19]. What the missing data problem adds is the component that the observation mechanism inflates, together with its part in retention, in the envelope and in the phase assignment. Ma et al. [20] study worst-case estimation over contamination neighborhoods of data missing completely

at random (MCAR) that contain mechanisms not at random, where bias dominates the minimax error. The envelopes here stay inside MAR, so the estimand remains identified and precision is the only quantity left uncertain.

The article establishes the following. Theorems 1 and 2 prove an impossibility result at a fixed complete-data law: unless every admissible mechanism observes each unit with at least some fixed probability, no observation rate guarantees that a nontrivial share of the information survives. Theorems 3 to 5 derive the exact favorable and adverse endpoints, in atomless and in mixed covariate spaces, and show that every precision between them is attained. Theorem 6 establishes the phase geometry described above. Corollaries 1 and 2 convert a relative or an absolute requirement into a single budget on the inflated component and, with the planning inputs held fixed, into the least costly set of additional observations. Theorems 7 to 9 supply cross-fitted inference on a prespecified grid of rates, conditions under which the refitted bootstrap remains valid, and the inversion that turns the resulting bands into an ordered confidence set for the two thresholds. Theorem 8 assumes a bootstrap expansion for the refit on a regular branch and transports that expansion to the implemented procedure. The theorem therefore states what the assumption delivers rather than verifying the assumption for any particular learner (Supplementary Section S2.4). Exact calculations, two illustrations on public clinical data, and a Monte Carlo study examine how the procedure behaves; the Supplementary Materials hold the proofs, the regularity conditions, and the extended computations.

2. Estimand-Specific Information Under MAR

2.1. Data Structure and Assumptions

Let $Z = (W,X)$, where W is always observed and X denotes one variable or one all-or-none block. The indicator R records whether the block was observed, so the observed data are $O = (W,R,RX)$. The scalar target $\theta = \Psi(P_Z)$ is pathwise differentiable in the locally nonparametric complete-data model, with canonical gradient φ_θ and finite second moment, and we observe independent and identically distributed copies $O_1,\ldots,O_n$. Patterns that leave a block partly observed and clustered or otherwise dependent sampling fall outside the present scope. Table 1 collects the notation used throughout.

Table 1. Notation. Subscripts identify the estimand or the index; superscripts carry the decorations defined in the last three rows. The letter C denotes a scalar when it carries the subscript θ and a set when θ appears as an argument.

Symbol	Meaning
θ	Scalar target parameter, $\theta = \Psi(P_Z)$.
$Z = (W, X)$	Complete data. W is always observed; X is the single variable or all-or-none block subject to missingness.
R, O	Observation indicator for X, and the observed data $O = (W, R, RX)$.
$\pi(W)$	Observation probability $P(R = 1 \mid W)$.
π_L	Lower bound imposed on $\pi(W)$ across the admissible class.
$\Pi(p,\pi_L)$	Admissible mechanisms with overall observation rate p and floor π_L.
$\varphi_\theta(Z)$	Complete-data efficient influence function for θ.
$a_\theta(W)$, $b_\theta(W)$	Conditional mean and conditional variance of $\varphi_\theta(Z)$ given W.
A_θ, B_θ	$\mathrm{Var}\{a_\theta(W)\}$ and $E\{b_\theta(W)\}$, the two complete-data variance components.
$C_\theta(\pi)$	$E\{b_\theta(W)/\pi(W)\}$: the residual component under one mechanism π. A scalar.
$V_{full}(\theta)$, $V_{obs}(\theta)$	Complete-data and observed-data efficiency bounds, $A_\theta + B_\theta$ and $A_\theta + C_\theta(\pi)$.

Table 1. *Cont.*

Symbol	Meaning
ρ_θ	Information retention $V_{full}(\theta)/V_{obs}(\theta)$.
$\gamma^{EB}{}_\theta$	Efficiency-bound fraction of information lost, $1 - \rho_\theta$.
$R^2(\theta)$	Share of $Var\{\varphi_\theta(Z)\}$ explained by W, $A_\theta/(A_\theta + B_\theta)$.
p, q	Overall observation rate and overall missingness rate, $q = 1 - p$.
q_{max}	Upper end of the admissible domain for the missingness rate, $q_{max} = 1 - \pi_L$. The inferential grid of Section 4.4 satisfies $q_L \leq q_{max}$ but need not reach it.
$C^-(\theta,p)$, $C^+(\theta,p)$	Smallest and largest value of $C_\theta(\pi)$ over $\Pi(p,\pi_L)$: the envelope endpoints.
$C(\theta,p)$	The attainable set $\{C_\theta(\pi): \pi \in \Pi(p,\pi_L)\}$, which equals $[C^-(\theta,p), C^+(\theta,p)]$. A set, not a scalar.
C_{req}	Residual budget the precision requirement allows. Equal to $C_0(\theta)$ or to $C_{lim}(\theta)$.
$C_0(\theta)$	C_{req} for a relative retention requirement at level ρ_0.
V_{lim}, $C_{lim}(\theta)$	Prespecified per-observation variance limit and the matching $C_{req} = V_{lim} - A_\theta$.
$\Delta_G(p)$, $\Delta_P(p)$	Guaranteed and possible margins, $C_{req} - C^+(p)$ and $C_{req} - C^-(p)$.
$q_G{}^*$, $q_P{}^*$	Guaranteed and possible thresholds on the overall missingness rate.
$q_{MCAR}{}^*$	Corresponding threshold when the mechanism is MCAR.
q_{Eq}	MCAR-equivalent missingness rate: the MCAR rate giving the same retention.
$\lambda(p)$, λ^*	Lagrange multipliers for the favorable-allocation and the minimum-cost programs.
$L_{k\ell}$, $U_{k\ell}$	Lower and upper simultaneous confidence limits for the margin Δ_k at grid rate q_ℓ.
$\mathcal{F}$	A sub-σ-field of σ(W): the resolution at which missingness may select units (Proposition S1).
Subscript lim	A prespecified target or limit, as in V_{lim} and C_{lim}.
Superscript *	A threshold on q, as in $q_G{}^*$; an optimum, as in $e_1{}^*$; or a bootstrap quantity, as in $\hat{\theta}^*$.
Hat	Sample estimate of the quantity carried under it, as in $\hat{C}^+(\theta,p)$ and $\hat{\rho}$.

For target θ define

$$a_\theta(W) = E\{\varphi_\theta(Z) \mid W\},\ b_\theta(W) = \mathrm{Var}\{\varphi_\theta(Z) \mid W\} \tag{1}$$

The function $b_\theta(W)$ depends on the estimand. It measures what a unit with covariates W still contributes to the uncertainty of θ once the always observed information has been used.

Assumption 1 (MAR). R is conditionally independent of X given W; equivalently, $P(R = 1 \mid X,W) = P(R = 1 \mid W)$ almost surely.

Assumption 2 (estimand-level overlap and integrability). The observation probability $\pi(W) = P(R = 1 \mid W)$ is positive almost surely and $C_\theta(\pi) = E\{b_\theta(W)/\pi(W)\}$ is finite.

Assumption 3 (locally nonparametric MAR reference model). At P_Z the closure of the complete-data tangent space is $L_2{}^0(P_Z)$, and the observation mechanism is variation independent of P_Z and otherwise unrestricted subject to Assumptions 1–2. The observed-data tangent closure then contains the three families of directions used in the coarsening at random projection, namely those for the marginal law of W, for the conditional law of X given W, and for the propensity score.

Every envelope below is defined within this locally nonparametric model. Under semiparametric restrictions the efficiency representation of Theorem 1 remains a benchmark, although it need not be the canonical gradient of the restricted model [9,21]. Theorems 2 to 6

then transfer only when the restricted bound keeps the separable form $A_\theta + E\{b_\theta(W)/\pi(W)\}$, with b_θ not depending on π.

2.2. Efficiency Bounds and Information Retention

Theorem 1 (locally nonparametric MAR efficiency representation). Under Assumptions 1–3, the observed-data efficient influence function is

$$D_\theta(O) = a_\theta(W) + R/\pi(W) \cdot [\varphi_\theta(Z) - a_\theta(W)] \tag{2}$$

$$A_\theta = \mathrm{Var}\{a_\theta(W)\},\ B_\theta = E\{b_\theta(W)\},\ C_\theta(\pi) = E\{b_\theta(W)/\pi(W)\} \tag{3}$$

$$V_{full}(\theta) = A_\theta + B_\theta,\ V_{obs}(\theta) = A_\theta + C_\theta(\pi) \tag{4}$$

Supplementary Section S1.1 verifies the score identity, the orthogonality to nuisance directions, and membership in the tangent space. The conditional variance decomposition then gives (4), which separates the information that W predicts from the residual influence variation, the latter inflated by inverse observation probabilities.

$$\rho_\theta = V_{full}(\theta)/V_{obs}(\theta) = (A_\theta + B_\theta)/\{A_\theta + C_\theta(\pi)\} \tag{5}$$

We call ρ_θ the estimand-specific information retention ratio. It is a ratio of variances, not a correlation. The following three quantities ensue directly from it: the standard error ratio $1/\sqrt{\rho_\theta}$, the complete-data-equivalent sample size $n\rho_\theta$, and the sample size multiplier $1/\rho_\theta$.

To keep this population quantity distinct from what a particular imputation procedure reports, define the efficiency-bound fraction of information lost as

$$\gamma^{EB}{}_\theta = 1 - \rho_\theta = 1 - V_{full}(\theta)/V_{obs}(\theta) \tag{5a}$$

This loss depends on the complete-data law, on the MAR mechanism, and on the estimand. It describes the efficiency bound rather than an imputation procedure, so the number of imputations does not enter [22].

2.3. MCAR-Equivalent Representation and Measures of Missing Information

For $B_\theta > 0$ define the observation rate of the MCAR mechanism with the same efficiency bound:

$$p_{Eq}(\theta) = B_\theta / C_\theta(\pi),\ q_{Eq}(\theta) = 1 - p_{Eq}(\theta) \tag{6}$$

If $B_\theta = 0$, q_{Eq} is undefined but $\rho_\theta = 1$. Otherwise, with $R^2(\theta) = A_\theta / V_{full}(\theta)$, the actual MAR mechanism has the same precision as MCAR at p_{Eq}, and

$$\rho_\theta = (1 - q_{Eq}(\theta))/(1 - R^2(\theta)\, q_{Eq}(\theta)) \tag{7}$$

The equivalent rate q_{Eq}, the efficiency-bound loss $\gamma^{EB}{}_\theta$, and Rubin-type fractions of missing information differ from one another. Supplementary Section S1.2 and Table S1 give the exact relations among them and show how far the correspondence with an ideal imputation extends [23–25].

An exact decomposition separates variation in the observation probability from the alignment between residual influence and missingness:

$$1/p_{Eq}(\theta) = 1/p + [E\{1/\pi(W)\} - 1/p] + \mathrm{Cov}\{b_\theta(W), 1/\pi(W)\}/B_\theta \tag{8}$$

The middle term never falls below zero, and the covariance term turns adverse when units with larger residual influence variance are observed less often.

2.4. No Universal Cutoff on the Overall Rate

Theorem 2 (no universal guarantee from the overall rate without uniform overlapping). Consider the mean estimand in the locally nonparametric MAR model and require each observation mechanism to be strictly positive but do not impose positive lower bound shared by the mechanism class. (i) There exists one square-integrable complete-data law P_0, independent of q and κ, such that for every overall missingness rate $q \in (0, 1)$ and every $\kappa \in (0, 1)$, an individually strictly positive MAR mechanism π_A exists with $E\{1 - \pi_A(W)\} = q$ and $\rho_\theta(\pi_A) < \kappa$. (ii) Moreover, for every $q \in (0, 1)$ and $\kappa \in (0, 1)$, there exist one square-integrable complete-data law $P_{q,\kappa}$ and two individually strictly positive MAR mechanisms π_F and π_A defined under that same law, each with $E\{1 - \pi_j(W)\} = q$, such that $\rho_\theta(\pi_F) = 1$ and $\rho_\theta(\pi_A) < \kappa$.

Part (i) fixes one complete-data law for all q and κ, while part (ii) fixes the law within each pair of favorable and adverse mechanisms. Positivity of each individual mechanism therefore guarantees nothing about precision from the observation rate alone. Supplementary Section S1.3 proves the theorem and compares the two quantifier orders in detail.

The construction moves the familiar difficulty of observation probabilities close to zero from a single mechanism to a class of them [26]: every mechanism here stays strictly positive, and only a floor shared by the whole class fails to exist.

3. Sharp Precision Envelopes at a Fixed Overall Rate

3.1. Admissible Mechanisms

Fix a lower bound $\pi_L \in (0, 1)$ on the observation probability and an overall rate $p \in [\pi_L, 1]$. This bound defines the mechanism class and need not equal the smallest propensity estimated in the realized dataset. Define

$$\Pi(p, \pi_L) = \{\pi: \pi_L \leq \pi(W) \leq 1 \text{ and } E\{\pi(W)\} = p\} \tag{9}$$

$$C^-(\theta,p) = \inf\{E[b_\theta(W)/\pi(W)]: \pi \in \Pi(p,\pi_L)\} \tag{10a}$$

$$C^+(\theta,p) = \sup\{E[b_\theta(W)/\pi(W)]: \pi \in \Pi(p,\pi_L)\} \tag{10b}$$

$C^-(\theta,p)$ corresponds to the most favorable admissible allocation of missingness and $C^+(\theta,p)$ to the most adverse.

3.2. Best Admissible Allocation

Theorem 3 (best allocation). Work under the envelope class in (9), with $0 < \pi_L < 1$, $b_\theta(W)$ measurable and nonnegative, and $B_\theta = E\{b_\theta(W)\} < \infty$. Let $r_\theta = P\{b_\theta(W) > 0\}$ and $p_{Sat}(\theta) = \pi_L + (1 - \pi_L)r_\theta$. If $B_\theta > 0$ and $\pi_L < p < p_{Sat}(\theta)$, the unique optimal allocation (up to null sets) is

$$\pi_{best}(p,W) = \text{clip}\{\text{sqrt}[b_\theta(W)/\lambda(p)], \pi_L, 1\} \tag{11}$$

The multiplier $\lambda(p) > 0$ is chosen so that $E\{\pi_{best}\} = p$. At $p = \pi_L$ only the constant allocation π_L is admissible; at $p = p_{Sat}(\theta)$ the optimum observes $\{b_\theta > 0\}$ completely and uses π_L on $\{b_\theta = 0\}$; above $p_{Sat}(\theta)$ every minimizer equals one on $\{b_\theta > 0\}$; and if $B_\theta = 0$ every allocation is optimal. The clipped square root in (11) is box-constrained Neyman allocation, and Supplementary Section S1.4 gives the optimizer sets and the boundary proof.

3.3. Worst Admissible Allocation

Theorem 4 (worst allocation in an atomless covariate space). Work under the envelope class in (9), with $0 < \pi_L < 1$, $b_\theta(W)$ measurable and nonnegative, $B_\theta = E\{b_\theta(W)\} < \infty$, and

$p \in [\pi_L,1]$. Suppose the probability space generated by W is atomless. Let $a(p) = (1 - p)/(1 - \pi_L)$ and let $Q_b(u)$ be any generalized quantile function of $b_\theta(W)$. Then

$$C^+(\theta,p) = B_\theta + (1/\pi_L - 1) \int [1 - a(p), 1]\, Q_b(u)\, du \tag{12}$$

The endpoint is explicit rather than the value of a numerical program. Because it integrates the quantile function of b_θ, a single sorting of the estimated residual influence variances evaluates it at every rate. For $0 < a(p) < 1$ the maximizers place π_L on an upper tail of b_θ of mass $a(p)$, splitting any tie at a positive threshold in a measurable way. At the two ends of the range the maximizers are $\pi \equiv 1$ and $\pi \equiv \pi_L$, and when $B_\theta = 0$ every allocation is maximizing.

In an atomless space the upper-tail bound is attained for every integrable b_θ and every rate, with no density assumption, whereas a single atom can leave it unattained. Supplementary Section S1.5 gives the maximizing set and a counterexample.

3.4. General Mixed Spaces and the Complete Attainable Set

Theorem 5 (adverse envelope on a general mixed space and completeness of the attainable set). Decompose the probability space generated by W into an atomless component N of mass w_0 and at most countably many atoms A_j of masses w_j, on which $b_\theta(W) = b_{\theta,j}$ almost surely. Put $c = 1 - \pi_L$, $a = (1 - p)/c$, $k = 1/\pi_L - 1$, and $H_N(s) = \sup\{\int_S b_\theta\, dP : S \subset N, P(S) = s\}$. Then, the adverse endpoint is the attained maximum

$$C^+(\theta,p) = B_\theta + \max\,[k\, H_N(s) + \Sigma_j\, w_j\, b_{\theta,j}\, \{1/[1 - c\, x_j] - 1\}], \tag{12a}$$

where the maximum is over $0 \le s \le w_0$ and $0 \le x_j \le 1$ satisfying

$$s + \Sigma_j\, w_j\, x_j = a. \tag{12b}$$

The atomless reduction uses the bathtub principle, and the atomic maximizer uses the Bauer maximum principle [27–29]. Supplementary Section S1.6 gives proof and shows that at most one atom is split.

Formula (12a) reduces to (12) when the space has no atoms and to exact vertex enumeration when it has finitely many strata. The accompanying software solves the finite atomic case, and the regularity conditions used for inference rule out unstable switching between atomic optimizers. Collect the attainable values of the residual component as

$$C(\theta,p) = \{E[b_\theta(W)/\pi(W)] : \pi \in \Pi(p,\pi_L)\}. \tag{12c}$$

Then, $C(\theta,p) = [C^-(\theta,p), C^+(\theta,p)]$, so the information retention set is also a closed interval when $V_{full}(\theta) > 0$: the envelope contains every attainable precision rather than only its two endpoints.

Suppose the class is restricted to mechanisms measurable with respect to a coarser σ-field $\mathcal{F} \subseteq \sigma(W)$, which represents the resolution at which missingness may select units. The envelopes then nest monotonically in $\mathcal{F}$ and reduce to the formulas above with b_θ replaced by $E\{b_\theta(W) \mid \mathcal{F}\}$ (Proposition S1, Supplementary Section S1.6). Coarsening that resolution therefore narrows the envelope at the same overall rate. The rise in q_G^* from 5.3% to 17.2% when allocation is restricted to the two stratum labels, reported in Section 5.1, is one instance, and resolution joins targeted recovery (Corollary 2) as a second quantity the design can control.

3.5. Precision Phases and Thresholds on the Overall Rate

Let $\rho_0 \in (0, 1)$ be the minimum information retention ratio the study requires. An equivalent form of the requirement uses the maximum admissible residual component

$$C_0(\theta) = V_{full}(\theta)/\rho_0 - A_\theta \quad (13)$$

The two crossing rates of the requirement are then

$$q_G^*(\theta) = \sup\{q \in [0, 1 - \pi_L]: C^+(\theta,1 - q) \leq C_0(\theta)\} \quad (14)$$

$$q_P^*(\theta) = \sup\{q \in [0, 1 - \pi_L]: C^-(\theta,1 - q) \leq C_0(\theta)\} \quad (15)$$

Theorem 6 (global phase geometry). If $V_{full}(\theta) > 0$, both endpoints are finite, continuous, and nonincreasing in $p \in [\pi_L, 1]$, with common values B_θ at $p = 1$ and B_θ/π_L at $p = \pi_L$. The acceptance sets are closed initial intervals $[0, q_G^*(\theta)]$ and $[0, q_P^*(\theta)]$, where $0 < q_G^*(\theta) \leq q_P^*(\theta) \leq 1 - \pi_L$, giving

guaranteed: $0 \leq q \leq q_G^*$; allocation-sensitive: $q_G^* < q \leq q_P^*$; and inadmissible: $q_P^* < q \leq 1 - \pi_L$, (15a)

Equality and plateaus are assigned to the more favorable phase. The allocation-sensitive region is empty exactly when $q_G^* = q_P^*$, and the inadmissible region is empty exactly when $q_P^* = 1 - \pi_L$ (Supplementary Section S1.7). Continuity follows from standard maximum theorem arguments applied to both exact endpoints, which is what yields a single global phase map. Table 2 sets out the three phases and the condition that defines each.

Table 2. Estimand-specific precision phases at observation rate p for the selected residual budget C_{req}.

Phase	Mathematical Condition	Interpretation
Guaranteed precision-admissible	$C^+(\theta,p) \leq C_{req}(\theta)$	Every admissible allocation meets the selected relative or absolute precision requirement.
Allocation-sensitive	$C^-(\theta,p) \leq C_{req}(\theta) < C^+(\theta,p)$	Whether the selected requirement is met depends on which units are missing.
Guaranteed precision-inadmissible	$C^-(\theta,p) > C_{req}(\theta)$	Even the most favorable allocation fails the selected precision requirement.

Under MCAR, $\rho_{MCAR}(q) = (1 - q)/\{1 - R^2(\theta)q\}$, and the crossing rate, truncated to the admissible domain, is

$$q_{MCAR}^*(\theta) = \min\{(1 - \rho_0)/(1 - \rho_0 R^2(\theta)), 1 - \pi_L\} \quad (16)$$

The MCAR crossing is a population benchmark only and lies between the sharp thresholds:

$$q_G^*(\theta) \leq q_{MCAR}^*(\theta) \leq q_P^*(\theta) \quad (17)$$

3.6. Absolute Precision Targets

Relative retention alone does not show whether the precision of a study is adequate. For a prespecified maximum per-observation variance V_{lim}, define $C_{lim}(\theta) = V_{lim} - A_\theta$. Then

$$V_{obs}(\theta) \leq V_{lim} \Leftrightarrow C_\theta(\pi) \leq C_{lim}(\theta). \quad (17a)$$

For a two-sided Wald interval with prespecified n, level $1 - \alpha$, and maximum half-width h_{lim}, and with $z(\cdot)$ the standard normal quantile function, the requirement translates into

$$V_{lim} = n\{h_{lim}/z(1 - \alpha/2)\}^2. \quad (17b)$$

Corollary 1 (phase geometry for an absolute target). If $V_{lim} > V_{full}(\theta)$, Theorem 6 applies with $\rho_{lim} = V_{full}(\theta)/V_{lim}$. If $V_{lim} = V_{full}(\theta)$, complete data meet the target only at equality. If $V_{lim} < V_{full}(\theta)$, the target cannot be met even with complete data, and the shortfall should not be attributed to missingness (boundary cases in Supplementary Section S1.8).

3.7. Minimum-Cost Targeted Recovery Under a Fixed Planning Law

Let $e_0(W) > 0$ be a fixed pilot observation law, $c(W) > 0$ a fixed marginal recovery cost, and C_{req} the residual budget the study has chosen. For $e_0 \leq e_1 \leq 1$, write the minimum recovery cost as

$$K^*(C_{req};e_0) = \inf E[c(W)\{e_1(W) - e_0(W)\}] \text{ subject to } E\{b_\theta(W)/e_1(W)\} \leq C_{req}. \quad (17c)$$

Corollary 2 (minimum-cost targeted recovery). No recovery is needed when $C_\theta(e_0) \leq C_{req}$, whereas $C_{req} < B_\theta$ is infeasible. For $B_\theta < C_{req} < C_\theta(e_0)$, every optimum has the following cost-adjusted square-root form, with λ^* chosen to make the constraint bind; at $C_{req} = B_\theta$, all units with $b_\theta > 0$ require recovery. The proof is in Supplementary Section S1.9.

$$e_1{}^*(W) = \min [1, \max\{e_0(W), \sqrt{\{\lambda^* b_\theta(W)/c(W)\}}\}]. \quad (17d)$$

$$s^*(W) = \{e_1{}^*(W) - e_0(W)\}/\{1 - e_0(W)\}. \quad (17e)$$

Recovering missing outcomes independently and conditionally uses the probability s^* of (17e), and the initial gain per unit cost is proportional to $b_\theta/\{ce_0{}^2\}$. The rule is cost-adjusted Neyman allocation under a box constraint, applied to the envelope budget and conditional on fixed e_0, b_θ, c, and C_{req}. Updating the allocation adaptively within the same study would require a separate multiwave or active design. Randomizing recovery conditionally on W preserves MAR, which recovery left uncontrolled need not do.

4. Estimation for a Partially Observed Continuous Outcome

4.1. Nuisance Functions and Variance Components

Let O = (W, R, RY), where W is fully observed and the continuous outcome Y may be missing. Under MAR, define $e(W) = P(R = 1 \mid W)$, $m(W) = E(Y \mid W)$, and $v(W) = Var(Y \mid W)$. In the notation of Section 2, the fully observed part plays the role of W, and the outcome Y plays the role of the block that may be missing. For $\theta = E(Y)$, the complete-data influence function is $Y - \theta$, so

$$A = Var\{m(W)\}, B = E\{v(W)\}, C = E\{v(W)/e(W)\} \quad (18)$$

Here, A, B, and C are A_θ, B_θ, and $C_\theta(\pi)$ for this estimand, and P_n denotes the empirical measure. Cross-fitting supplies out-of-fold predictions $\hat{m}i$, $\hat{v}i$, and $\hat{e}i$ from training data that exclude observation i [30]. The one-step estimators are then

$$\hat{\mu} = P_n[\hat{m} + R/\hat{e} \cdot (Y - \hat{m})] \quad (19)$$

$$\hat{A} = P_n[(\hat{m} - \hat{\mu})^2 + 2R/\hat{e} \cdot (\hat{m} - \hat{\mu})(Y - \hat{m})] \quad (20)$$

$$\hat{B} = P_n[\hat{v} + R/\hat{e} \cdot \{(Y - \hat{m})^2 - \hat{v}\}] \quad (21)$$

$$\hat{C} = P_n[v^{\wedge}/\hat{e} + R/\hat{e}^2 \cdot \{(Y - m^{\wedge})^2 - v^{\wedge}\} - v^{\wedge}/\hat{e}^2 \cdot (R - \hat{e})] \quad (22)$$

The estimated bounds are $\hat{A} + B^{\wedge}$ and $\hat{A} + \hat{C}$. The one-step forms propagate uncertainty in both the outcome model and the observation model; Supplementary Section S2.1 gives the influence functions and the remainder conditions. Theorem 7 requires convergence rates for all three nuisance functions. Substituting a flexible nonparametric learner does not remove that requirement; the rate has to be established for the learner in use, and the implementation does not verify that rate. Section 6.4 reports joint coverage of 54.0% under a random forest whose convergence rate was not established.

4.2. Envelope Estimation

At each prespecified p, the empirical analog of the square root rule (11) gives $\hat{C}^-$ below the estimated saturation point, and the upper-tail rule of (12) gives $\hat{C}^+$. Boundary allocations apply at saturation and at the ends of the domain.

The empirical favorable endpoint is the plug-in optimizer in one-step form, which is the estimator validated in the nine-cell study of Section 6.3. Other admissible allocations can attain a smaller estimated residual component, the constant allocation $\pi \equiv p$ among them, and taking the smaller of two such allocations is itself admissible. A minimum of this kind becomes nonregular at ties, however, and would need validation of its own (Supplementary Section S2.6). Supplementary Section S5.8 reports a paired examination of two such rules, neither of which is adopted here. The primary algorithm is the plug-in optimizer alone, and it produced every estimate reported in this article.

Because the clipping and tail thresholds are themselves estimated, the conditions of Supplementary Section S2.3 control the margin around them and the displacement of the boundary, and these conditions go beyond marginal L_2 rates for the nuisance estimators. Alerts raised at a boundary or at saturation are descriptive. Write the guaranteed margin and the possible margin against the requirement as

$$\Delta_G(p) = C_{req} - C^+(p), \Delta_P(p) = C_{req} - C^-(p) \quad (23)$$

Here, C_{req} is C_0 for a relative retention requirement and C_{lim} for an absolute one. The phase is guaranteed when $\Delta_G \geq 0$, allocation-sensitive when $\Delta_G < 0 \leq \Delta_P$, and inadmissible when $\Delta_P < 0$, with equality assigned to the admissible side.

4.3. Simultaneous Inference on a Prespecified Finite Grid

Let $G = \{p_1, \ldots, p_L\}$ be fixed before fitting. The quantity estimated for inference stacks the 2 L phase margins and, when the requirement is absolute, the feasibility margin $F_{lim} = V_{lim} - (A + B)$; no phase is certified unless feasibility itself is established. The guarantees cover this fixed grid alone. They do not extend to a continuum band, to a grid chosen from the data, or to an interval for a smooth crossing.

Theorem 7 (cross-fitted fixed-grid expansion). Under conditions R1–R6 of Supplementary Section S2.3, the fixed-dimensional estimator is jointly asymptotically linear, and its studentized maximum has a Gaussian limit that the centered multiplier approximates consistently. Cross-fitting does not replace the requirements of the optimizer and boundary displacement.

Theorem 8 (full-refit bootstrap validity). Under Theorem 7 and conditions B1–B6 of Supplementary Section S2.3, the ideal implemented law is consistent when the one-refit failure probability $q_n^* = o_P(1)$, and the fail-closed empirical critical value has simultaneous coverage $1 - \alpha + o(1)$ when $B_n \to \infty$ and $B_n q_n^* = o_P(1)$; fixed B adds Monte Carlo error.

The recommended procedure refits every nuisance model and rerandomizes all splits. It uses the kth maximum studentized deviation with $k = \text{ceil}\{(Bs + 1)(1 - \alpha)\}$, which gives

$k = 950$ for $Bs = 999$ and $\alpha = 0.05$, centers the intervals at the original estimate, and fails closed on any unusable refit. The plus-one convention does not deliver exact finite-sample coverage. Condition B3 is an assumption in its own right. Supplementary Section S2.4 gives a counterexample showing that it does not follow from the rate and success conditions, and verifying B3 formally for general machine learners remains an open question, so the coverage statement of Theorem 8 holds conditionally on B3 for the learner in use.

A phase is certified only when the simultaneous intervals strictly establish the signs that define it. Boundary cases and all others are reported as indeterminate, and the feasibility coordinate is treated in the same way. On the event that the intervals cover jointly, every statement other than indeterminate is correct.

4.4. Confidence Set for the Global Thresholds by Band Inversion

Write the missingness grid as $0 < q_1 < \ldots < q_L \leq q_{max}$, where $q_\ell = 1 - p_\ell$ and $q_{max} = 1 - \pi_L$ is the upper end of the admissible domain defined in Section 3.1, not the largest rate on the grid. The grid need not reach q_{max}, so a threshold can lie beyond q_L; the anchors in (25) are set at q_{max} for that reason. Let $L_{k\ell}$ and $U_{k\ell}$ denote the lower and upper simultaneous confidence limits of Section 4.3 for the margin Δ_k at rate q_ℓ. For margin $k \in \{G, P\}$, monotonicity permits the coverage-preserving tightening

$$\tilde{L}_{kj} = \max(\ell \geq j)\, L_{k\ell},\ \tilde{U}_{kj} = \min(\ell \leq j)\, U_{k\ell}. \tag{24}$$

Define strict, fail-closed threshold brackets, for $k \in \{G, P\}$,

$$q_k^{lo} = \max(\{0\} \cup \{q_j : \tilde{L}_{kj} > 0\}),\ q_k^{up} = \min(\{q_{max}\} \cup \{q_j : \tilde{U}_{kj} < 0\}). \tag{25}$$

The ordered joint confidence set is

$$Q_{1-\alpha} = ([q_G^{lo}, q_G^{up}] \times [q_P^{lo}, q_P^{up}]) \cap \{(q_G, q_P) : q_G \leq q_P\}. \tag{26}$$

Theorem 9 (ordered threshold confidence set). If the simultaneous margins cover with probability $1 - \alpha + o(1)$, $Q_{1-\alpha}$ contains (q_G^*, q_P^*) jointly with at least that probability. On the same event, rates below q_G^{lo} are correctly certified as guaranteed, rates between q_G^{up} and q_P^{lo} as allocation-sensitive, and rates above q_P^{up} as inadmissible, while all other rates are indeterminate; conclusions about an absolute target additionally require certified complete-data feasibility. The construction is standard confidence set inversion [31], it is limited by the fixed grid, and it fails closed if the tightening or the order intersection is empty. Because $q_{max} = 1 - \pi_L$, an upper limit returned at that anchor records that no grid rate excluded the threshold, not that the threshold equals the largest rate examined.

4.5. Relation to Multiple Imputation

The envelope does not depend on the method used for analysis. If the intrinsic bound is not met, no regular implementation can attain the target, including multiple imputation that is efficient and correctly specified. Meeting the bound shows only that sufficient information remains; specification, congeniality, the number of imputations, and learnability still govern what an analysis achieves.

4.6. Implementation

The implementation accepts a CSV file holding a partially observed continuous outcome, its binary observation indicator, and completely observed numeric covariates. It validates the structure of the file, cross-fits the nuisance functions, estimates realized retention and both envelopes, runs the fail-closed full-refit bootstrap, certifies phases, inverts the threshold sets, and records diagnostics. A descriptive recovery plan under fixed inputs

is optional. Every output is machine-readable and comes with a run manifest. Two limits apply to the reported inference. The bootstrap output is conditional on B3 for the nuisance learners actually used, and the implementation does not check that assumption. The simultaneous level is also nominal rather than guaranteed: Section 6 records joint coverage below 95% at $n = 500$ and in the boundary and random-forest configurations, so no blanket 95% claim attaches to the output at small sample sizes or near the boundary. The same qualification applies more strongly to the fast multiplier procedure, which stays further from nominal throughout, with joint coverage between 76.8% and 87.4% in the Monte Carlo study of Section 6; it is reported as a diagnostic and not as a basis for a simultaneous claim.

5. Illustrations

5.1. Exact Population Calculation

Consider a synthetic imaging study of a mean continuous biomarker Y that is observed only after an optional physiologic acquisition. The baseline W is complete, observation is MAR given W, A = B = 1, and the residual influence variance takes the values 0.1 and 1.9 on two regions of equal mass in an atomless space for W. No empirical patient data enter this example.

At 75% observation, the favorable allocation retains 97.6% of the complete-data information, MCAR retains 85.7%, and the adverse allocation retains 67.8%. For $n = 1000$ these correspond to 976, 857, and 678 complete-data-equivalent observations, and to standard error increases of 1.2%, 8.0%, and 21.4%. The requirement $\rho_0 = 0.80$ is equivalent to $V_{lim} = 2.5$ and to a 95% half-width limit of 0.098. The favorable allocation and MCAR meet it; the adverse allocation does not. Starting from the adverse allocation, recovering 7.76 outcomes per 100 participants at equal cost raises observation to 82.76% and brings the study within the budget.

With $\pi_L = 0.20$, the exact thresholds are $q_G^* = 5.3\%$, $q_{MCAR}^* = 33.3\%$, and $q_P^* = 52.0\%$. Restricting allocation to the two stratum labels raises q_G^* to 17.2%. Figure 1 and Tables S4 and S7 give the calculations and the sensitivity to π_L.

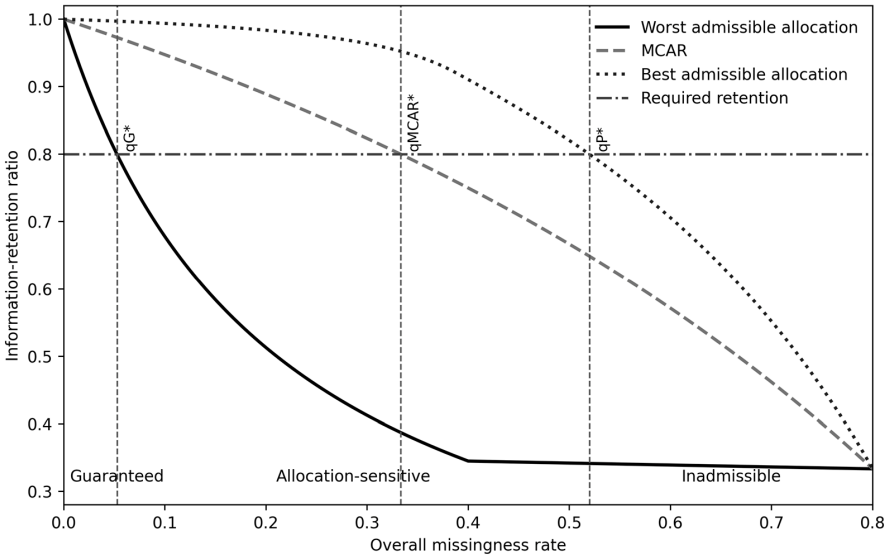

Figure 1. Exact sharp precision envelope for the analytic illustration. The lower curve is the worst admissible allocation and the upper curve is the best admissible allocation; MCAR lies between them. Vertical lines mark the population crossings q_G^*, q_{MCAR}^*, and q_P^*; they are not estimated confidence limits. The horizontal axis covers the whole admissible domain, 0 to $1 - \pi_L$ with $\pi_L = 0.20$. At its right endpoint the mean constraint leaves $\pi \equiv \pi_L$ as the only admissible allocation, so the three curves meet.

5.2. Fully Synthetic Analysis of Individual Records

The calculation above is exact because the law is fixed. The same procedure also runs on individual records, and the synthetic example below traces that path from a data file to a certified phase. The file holds 600 records and was generated without empirical patient or animal data. Each record carries two completely observed numeric covariates, a continuous biomarker analog with its binary observation indicator, and a positive recovery cost. The seed was fixed at 20260830. The outcome was observed in 448 records, an overall observation rate of 0.7467. The analysis fixed p = {0.60, 0.70, 0.75, 0.80, 0.90, and 0.97}, $\pi_L = 0.20$, and $\rho_0 = 0.80$, used two-fold outer cross-fitting with inner cross-residuals for variance learning, and ran B = 999 primary fail-closed refits.

The estimated variance bound under complete data was 1.2929 and the bound under the observed data was 1.7718, giving realized retention 0.7297 with marginal 95% Wald interval 0.6677–0.7917. All 999 refits succeeded and the rank-950 critical value was 3.0948. No boundary diagnostic was flagged. Direct inversion placed q_G^* in [0, 0.20] and q_P^* in [0.20, 0.40]. Table S10 reports the grid decisions. They rest on the primary max-deviation intervals rather than on marginal ones, so a phase is certified only when the relevant interval separates from zero. The run demonstrates one software path and is not a coverage experiment.

The descriptive recovery plan treated the cross-fitted variance, propensity, and cost predictions as fixed. It raised the mean observation probability from 0.7458 to 0.7627, adding 1.69 outcomes per 100 records and recovering 6.64% of the outcomes that were missing. Row-level probabilities and priorities are written to recovery_plan.csv, together with the machine-readable summaries and the run manifest listed in Supplementary Section S5.3. The plan holds the planning inputs fixed and is neither an adaptive acquisition rule nor a coverage statement. Because the variance, propensity, and cost predictions are themselves estimated, the stated gain is conditional on them. A miscalibrated model for the conditional variance moves the priority ordering and the number of observations the plan requires, as the ACTG 175 record in Supplementary Section S5.9 shows. The overlap floor does not enter this calculation; Section 3.7 conditions it on the pilot law, the residual influence variance and the cost.

5.3. Public MASLD Cohort Illustration

We then applied the same implementation, unchanged, to the public survival::nafld1 study [32,33], which is released as a 90% random sample with dates removed and arbitrary identifiers. The analysis was restricted to the 3514 cases of metabolic dysfunction-associated steatotic liver disease (MASLD), so that no independence assumption was needed for the matched controls. Baseline body mass index (BMI) was observed in 2989 cases (85.06%). The target was mean baseline BMI at diagnosis, and W comprised age, sex, and centered quadratic age.

The analysis fixed $\rho_0 = 0.80$, $\pi_L = 0.05$, p = {0.70, 0.75, 0.80, 0.85, 0.90, and 0.97}, two-fold cross-fitting, and B = 999 fail-closed refits. Estimated observation probabilities ranged from 0.812 to 0.945, with none below 0.05. That range is a compatibility diagnostic; it validates neither MAR nor the floor.

Estimated mean baseline BMI was 34.24 kg/m^2. Realized retention was 0.858, with marginal 95% Wald interval 0.846–0.870, an increase of 8.0% in the standard error and about 3014 complete-data-equivalent cases. The sharp interval at $p = 0.85$ was [0.177, 0.865]. The fitted mechanism met the 0.80 requirement, but the simultaneous phase was allocation-sensitive, because the overall rate alone did not guarantee the requirement across the class. All 999 refits succeeded, with critical value 3.336 and no boundary flag; Table 3, Figure 2, and Tables S11–S13 report the details. The allocation-sensitive verdict does not depend on

the chosen floor. With everything else held fixed, varying π_L over 0.02, 0.05, 0.10 and 0.20 moves the adverse endpoint at $p = 0.85$ from 0.079 to 0.473 while the favorable endpoint remains near 0.865, and the point phase stays allocation-sensitive at every floor (Table S12). The floor sets how far the adverse endpoint falls, not the phase reported for this cohort.

Table 3. Public MASLD-case analysis of mean baseline BMI under the fixed 80% information-retention requirement.

Cases/BMI Observed	Observed Rate	Realized ρ (Marginal 95%)	Sharp ρ Interval at $p = 0.85$	Simultaneous Phase	Successful Refits
3514/2989	0.851	0.858 (0.846–0.870)	[0.177, 0.865]	Allocation-sensitive	999/999

Note. Mean baseline BMI was 34.24 kg/m^2. The envelope uses $\pi_L = 0.05$. The realized-ρ interval is marginal; the phase uses the simultaneous max-deviation band over six rates. The dataset, estimand, and grid were selected before the final B = 999 run but after exploration of datasets and specifications, so this is an operational illustration rather than a confirmatory application. Fitted propensity diagnostics do not validate MAR or overlap.

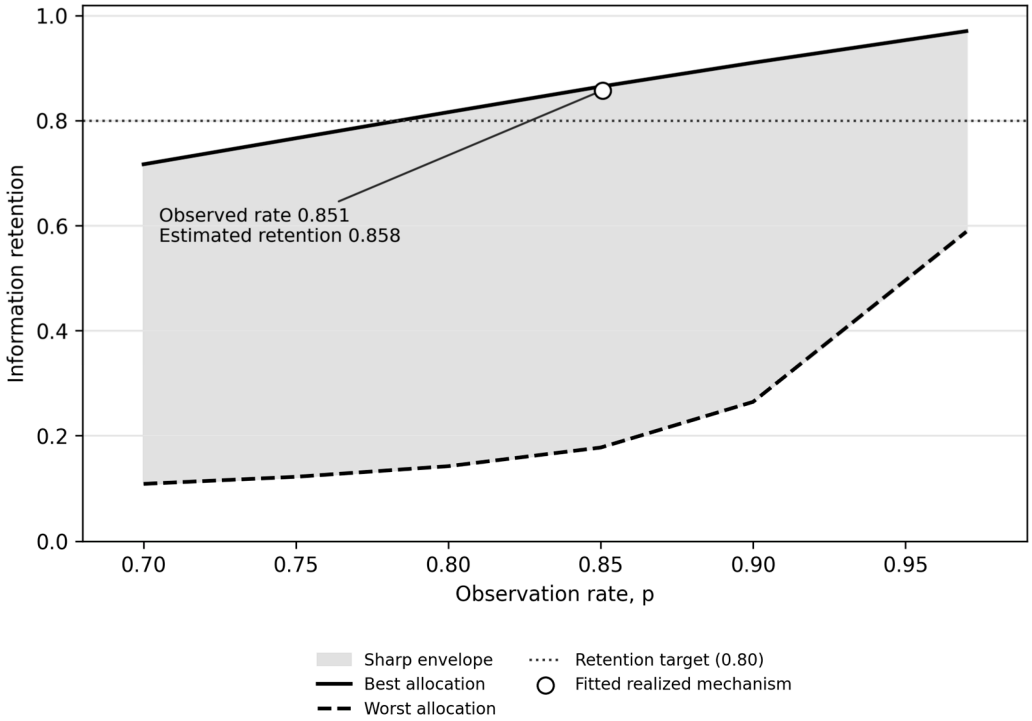

Figure 2. Public MASLD-case BMI illustration. The shaded region is the estimated sharp retention envelope under $\pi_L = 0.05$; the point is the fitted realized mechanism at the overall observation rate. The horizontal line is the fixed $\rho_0 = 0.80$ requirement.

A percentage rule would have reached a different conclusion here. The 14.9% of BMI values that were missing did not certify the 80% requirement, while the fitted mechanism met it. The appropriate report therefore gives both numbers rather than a declaration that 15% is safe.

A second public illustration, with 37% missingness and richer covariates, uses data from the AIDS Clinical Trials Group 175 trial (ACTG 175). Supplementary Sections S5.6 and S5.7 report that illustration together with the executed treatment-effect substitution of Supplementary Section S7. Supplementary Section S5.9 reports calibration diagnostics for the three nuisance models and repeats the analysis under a richer and a reduced covariate specification. In that illustration the fitted mechanism did not meet the 0.80 requirement

at the point estimate, with realized residual component 24,042 against a budget of 23,335, and that component fell below the estimated favorable endpoint of 26,698. Population endpoints cannot be ordered in that direction, so the reversal is a property of the fitted quantities rather than of the population: the calibration record shows the fitted conditional variance spreading its predictions over a sixfold range while the out-of-sample squared residuals vary by less than half as much, which leaves an estimated favorable allocation with almost no heterogeneity to exploit (Figure S3). Realized retention there was 0.785, with marginal 95% interval 0.752 to 0.819; that corresponds to an increase of 12.8% in the standard error and about 1680 complete-data-equivalent participants. No rate was certified as guaranteed or as allocation-sensitive at any of the three overlap floors examined. These quantities are therefore reported descriptively.

6. Monte Carlo Study

6.1. Design

Data were generated with W uniform on $(-1, 1)$, a linear conditional mean, conditional variance $v(W) = 1 + 0.8\,W$, and an average observation rate of 75%. The propensity slope took the values 1.5, 0, and -1.5, giving favorable, MCAR, and adverse mechanisms, at $n = 500$, 1000, and 2000. Fitting used two-fold cross-fitting, linear models for the mean and the variance, a logistic model for observation, and inner cross-residuals for variance fitting. The grid was $\rho = \{0.60, 0.70, 0.75, 0.80, 0.90, \text{and } 0.97\}$, with $\pi_L = 0.20$ and $\rho_0 = 0.80$.

Point and multiplier performance used 500 datasets per cell. The $B = 999$ validation used 1000 datasets in each of nine cells, with 999 attempted full refits each, or 8,991,000 refits in total. Any failed refit closed the outer dataset. Conditional coverage excludes such datasets, whereas operational coverage counts them as noncoverage.

6.2. Point Estimation and Fast Multiplier Simultaneous Intervals

Retention bias stayed below 0.005 and marginal coverage ran from 92.0% to 95.6%. Joint coverage from the fast multiplier procedure was much lower, between 76.8% and 87.4%, taken over all 12 grid coordinates, namely the guaranteed and possible margins at the six prespecified rates. The covariance-only approximation misses nuisance and curvature effects at these sample sizes. Table 4 reports the point-estimator performance and the joint coverage in full.

Table 4. Point-estimator performance and joint coverage across the prespecified finite grid for the fast centered influence-function Gaussian multiplier procedure (500 replicates per cell).

Mechanism	n	True ρ	Bias	Empirical SD	Mean SE	95% Coverage	Joint Grid Coverage
Favorable	500	0.838	+0.0010	0.0223	0.0200	93.2%	83.0%
Favorable	1000	0.838	+0.0001	0.0141	0.0134	93.0%	84.2%
Favorable	2000	0.838	+0.0005	0.0093	0.0093	95.0%	86.2%
MCAR	500	0.800	+0.0008	0.0219	0.0208	94.8%	76.8%
MCAR	1000	0.800	+0.0002	0.0155	0.0144	92.0%	86.6%
MCAR	2000	0.800	−0.0001	0.0103	0.0100	93.8%	86.0%
Adverse	500	0.714	+0.0045	0.0407	0.0380	93.4%	79.2%
Adverse	1000	0.714	−0.0016	0.0250	0.0255	95.6%	85.6%
Adverse	2000	0.714	+0.0007	0.0177	0.0173	94.4%	87.4%

6.3. Refitted Max-Deviation Bootstrap

Table 5 reports the $B = 999$ fail-closed validation. Of 9000 outer datasets, 8994 produced intervals and six failed closed. Conditional joint coverage ran from 93.1% to 95.4% and operational coverage from 92.7% to 95.3%. Only MCAR at $n = 500$ showed Holm-adjusted

undercoverage (p = 0.046), with adverse n = 500 borderline (0.053). No dataset falsely certified a phase, and the threshold sets obtained by inverting the labels contained both simulation-law thresholds in every successful dataset; that last check tests compatibility and is not independent evidence on coverage.

Table 5. B = 999 refitted max-deviation bootstrap performance under the primary fail-closed procedure (1000 attempted datasets per cell).

Mechanism	n	Successful/Attempted	Conditional Coverage (MCSE)	Operational Coverage	Mean Critical Value	Failed Datasets
Favorable	500	999/1000	95.4% (0.7%)	95.3%	3.08	1
Favorable	1000	1000/1000	95.0% (0.7%)	95.0%	2.98	0
Favorable	2000	1000/1000	93.8% (0.8%)	93.8%	2.92	0
MCAR	500	1000/1000	93.1% (0.8%)	93.1%	3.06	0
MCAR	1000	1000/1000	94.5% (0.7%)	94.5%	2.96	0
MCAR	2000	1000/1000	94.9% (0.7%)	94.9%	2.90	0
Adverse	500	995/1000	93.2% (0.8%)	92.7%	3.07	5
Adverse	1000	1000/1000	94.1% (0.7%)	94.1%	2.95	0
Adverse	2000	1000/1000	94.1% (0.7%)	94.1%	2.90	0

Note. Conditional coverage uses datasets that successfully produced an interval. Operational coverage counts a fail-closed outer dataset as noncoverage. MCSE is the Monte Carlo standard error of conditional coverage. Each dataset attempted 999 refits; six outer datasets failed closed after seven exception-type refit failures.

The estimator adopted in Section 4.2 is the plug-in optimizer. Two rules that also admit the constant allocation were examined against it and are not adopted, for the reasons given in Supplementary Section S5.8.

6.4. Structural and Boundary Stress Experiments

Across 200 randomized finite atomic spaces, the exact allocations met the mean constraint to within 10^{-12}, with no failure of monotonicity or of phase contiguity. In the smooth exact-boundary studies, joint coverage was 95.0% at the guaranteed boundary and 89.5% at the possible boundary, with false certification of 0.0% and 1.0% respectively. A correctly specified quadratic experiment gave 96.7% joint coverage. A random-forest experiment produced large endpoint errors and only 54.0% joint coverage from the fast procedure, which rules out a uniform finite-sample 95% claim and distinguishes the adequacy of nuisance learning from structural or finite-sample failure (Table S5).

7. Discussion

At the analysis stage the fitted retention says how much complete-data information survives for the estimand actually targeted, expressed as a standard error ratio and an equivalent sample size. The envelope then says how much of that conclusion rests on the realized allocation rather than on the rate. Before any mechanism exists, the same envelope delimits what percentage alone can support, with b_θ posited from pilot or external information rather than estimated.

Application requires the following four prespecified elements: the estimand, a relative or absolute precision target, a justified overlap floor or sensitivity grid, and a finite rate grid, with complete-data feasibility assessed first. When a target is missed but recoverable, the output is not a lower overall percentage but a priority ordering by $b_\theta/(ce_0^2)$ and the number of additional observations that ordering requires (Tables S8 and S9).

The B = 999 study supports the full-refit procedure but not uniform finite-sample exactness, so phase statements must carry their simultaneous margins, the computational success record, and Monte Carlo uncertainty. The bootstrap guarantee is also asymptotic

and conditional: Theorem 8 assumes the centered expansion B3 for the regular-branch refit, and Supplementary Section S2.4 gives a counterexample showing that this assumption does not follow from the rate and success conditions alone. Verifying that expansion for a given learner remains open. Nominal simultaneous coverage is accordingly not a blanket 95% guarantee at small n or near the boundary. The absence of false certifications recorded in Section 6.4 is a separate operating characteristic: a decision rule can avoid false certifications while the intervals behind it undercover, so both have to be reported side by side.

A post hoc $n = 500$ experiment found that no established refinement [34] improved coverage consistently, so the primary band was retained (Table S14).

Limitations are substantive. MAR and the overlap class are assumptions rather than diagnostics, and sensitivity to missingness that is not at random remains necessary. The public illustrations are retrospective and target descriptive means. Their nuisance models are the implemented low-dimensional ones, whose consistency for the data-generating law is not established; cross-fitting removes the bias that overfitting would introduce but leaves the effect of a misspecified functional form. Both illustrations therefore demonstrate the analysis path under a stated specification rather than deliver a confirmatory clinical conclusion, and neither justifies the overlap floor or a clinical threshold. Supplementary Section S5.9 reports how far the ACTG 175 quantities move when that specification changes. Observations are assumed independent and identically distributed, so clustered and longitudinal designs fall outside the present scope. The theory covers one variable or one all-or-none block, and the implementation covers one continuous outcome with complete numeric covariates; flexible learners require verification of their own rates. Regular inference excludes active boundary mass, saturation transitions, and unstable switching between atomic optimizers. The guarantee holds on the fixed grid alone, the recovery calculation is descriptive under fixed planning inputs, and the treatment-effect substitution of Supplementary Section S7 is executed only at the point level. Wider mechanism classes, among them odds ratio and divergence ball neighborhoods, multi-stage monotone missingness, and constraints on the cost of observation, are left to future work, and no conclusion above depends on them.

8. Conclusions

A missing-data percentage measures quantity, not the estimand-specific information that remains. Under MAR and a prespecified overlap class, the envelope reports what a study still knows about its own estimand, and whether that knowledge is guaranteed, dependent on the allocation, or unattainable at the observed rate. It also reports what a costed round of additional observation would have to achieve. Inference on the fixed grid then supports conservative phase statements and ordered threshold sets.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/stats9050110/s1, Supplementary Sections S1–S7, Tables S0–S16, and Figures S1–S3, together with the preparation scripts, analysis specifications, code, fixed simulation design and seeds, validation records, aggregate and replicate-level output, and SHA-256 checksums.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. Both illustrations are secondary methodological analyses of public, de-identified datasets distributed with the R packages survival and speff2trial. No participants were recruited, no identifiable records were accessed, and no intervention was performed. Under the Ethical Guidelines for Medical and Biological Research Involving Human Subjects (Public Notice No. 1 of 2021 of the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labor and Welfare, and the Ministry of Economy, Trade and Industry, Japan, as amended in 2023), studies that use only samples or information whose academic

value is already established, which are widely used for research and generally obtainable, fall outside the scope of the Guidelines (Chapter 1, Section 3), so no institutional review was required.

Informed Consent Statement: Informed consent for participation is not required as per local legislation: under the Guidelines cited above, studies that use only samples or information of that kind fall outside their scope. The study recruited no participants, and the individual records analyzed are de-identified as released.

Data Availability Statement: The survival::nafld1 cohort is distributed with the R survival package, with dates removed, identifiers arbitrary, and the file released as a 90% random sample for privacy; the institutional review approvals for the original cohort are reported by Allen et al. [33]. The ACTG 175 trial data are those distributed with the R speff2trial package. The preparation scripts, analysis specifications, code, fixed simulation design and seeds, validation records, aggregate and replicate-level output for all 9000 attempted outer datasets, and SHA-256 checksums are supplied as Supplementary Data; participant-level rows and row-level derived output are not duplicated there.

Acknowledgments: During the preparation of this manuscript, the author used OpenAI ChatGPT 5.6 for revising text and for prototyping code and analyses. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Groves, R.M.; Peytcheva, E. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opin. Q.* **2008**, *72*, 167–189. [CrossRef]
2. Madley-Dowd, P.; Hughes, R.; Tilling, K.; Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol.* **2019**, *110*, 63–73. [CrossRef] [PubMed]
3. Petersen, M.L.; Porter, K.E.; Gruber, S.; Wang, Y.; van der Laan, M.J. Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **2012**, *21*, 31–54. [CrossRef] [PubMed]
4. Junaid, K.P.; Kiran, T.; Gupta, M.; Kishore, K.; Siwatch, S. How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Popul. Health Metr.* **2025**, *23*, 2. [CrossRef] [PubMed]
5. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef] [PubMed]
6. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
7. Sterne, J.A.C.; White, I.R.; Carlin, J.B.; Spratt, M.; Royston, P.; Kenward, M.G.; Wood, A.M.; Carpenter, J.R. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* **2009**, *338*, b2393. [CrossRef] [PubMed]
8. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **1994**, *89*, 846–866. [CrossRef]
9. Graham, B.S. Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **2011**, *79*, 437–452. [CrossRef]
10. Neyman, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.* **1934**, *97*, 558–606. [CrossRef]
11. Gilbert, P.B.; Yu, X.; Rotnitzky, A. Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Stat. Med.* **2014**, *33*, 901–917. [CrossRef] [PubMed]
12. Tao, R.; Zeng, D.; Lin, D.-Y. Optimal designs of two-phase studies. *J. Am. Stat. Assoc.* **2020**, *115*, 1946–1959. [CrossRef] [PubMed]
13. Chen, T.; Lumley, T. Optimal sampling for design-based estimators of regression models. *Stat. Med.* **2022**, *41*, 1482–1497. [CrossRef] [PubMed]
14. Wesołowski, J.; Wieczorkowski, R.; Wójciak, W. Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata. *Surv. Methodol.* **2024**, *50*, 487–511.
15. Zrnic, T.; Candès, E. Active statistical inference. *Proc. Mach. Learn. Res.* **2024**, *235*, 62993–63010.
16. Ma, V.L.; Candès, E.J. Optimized labeling resource allocation for prediction-assisted inference via OPAL. *arXiv* **2026**, arXiv:2606.03211.
17. Wang, R.; Wang, Q.; Miao, W. A maximin optimal approach for sampling designs in two-phase studies. *arXiv* **2023**, arXiv:2312.10596.
18. Williamson, B.D.; Gilbert, P.B.; Carone, M.; Simon, N. Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **2021**, *77*, 9–22. [CrossRef] [PubMed]

19. Kish, L. *Survey Sampling*; Wiley: New York, NY, USA, 1965.
20. Ma, T.; Verchand, K.A.; Berrett, T.B.; Wang, T.; Samworth, R.J. Estimation beyond missing (completely) at random. *arXiv* **2024**, arXiv:2410.10704.
21. Robins, J.M.; Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* **1995**, *90*, 122–129. [CrossRef]
22. Louis, T.A. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 226–233. [CrossRef]
23. Barnard, J.; Rubin, D.B. Small-sample degrees of freedom with multiple imputation. *Biometrika* **1999**, *86*, 948–955. [CrossRef]
24. Meng, X.-L. Multiple-imputation inferences with uncongenial sources of input. *Stat. Sci.* **1994**, *9*, 538–558. [CrossRef]
25. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
26. Crump, R.K.; Hotz, V.J.; Imbens, G.W.; Mitnik, O.A. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **2009**, *96*, 187–199. [CrossRef]
27. Bauer, H. Minimalstellen von Funktionen und Extremalpunkte. *Arch. Math.* **1958**, *9*, 389–393. [CrossRef]
28. Lieb, E.H.; Loss, M. *Analysis*, 2nd ed.; American Mathematical Society: Providence, RI, USA, 2001. [CrossRef]
29. Aliprantis, C.D.; Border, K.C. *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2006. [CrossRef]
30. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **2018**, *21*, C1–C68. [CrossRef]
31. Neyman, J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. A* **1937**, *236*, 333–380. [CrossRef]
32. Therneau, T.M. Survival: A Package for Survival Analysis in R, Version 3.5-8. 2024. Available online: https://CRAN.R-project.org/package=survival (accessed on 1 September 2026).
33. Allen, A.M.; Therneau, T.M.; Larson, J.J.; Coward, A.; Somers, V.K.; Kamath, P.S. Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: A 20 year-community study. *Hepatology* **2018**, *67*, 1726–1736. [CrossRef] [PubMed]
34. Hall, P. Theoretical comparison of bootstrap confidence intervals. *Ann. Stat.* **1988**, *16*, 927–985. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.