Resource-Attributable Research Loss (RARL-R) in Postgraduate Medical Training: A Counterfactual Framework for Quantifying Remediable Research Opportunity Gaps

Zeshu Zhang[12*]

[1] Department of Neurosurgery, General Hospital of Ningxia Medical University, Yinchuan, Ningxia 750004, China

[2] Ningxia Medical University, Yinchuan, Ningxia 750004, China

* Corresponding author: Zeshu Zhang; Email: 240111120856@nxmu.edu.cn

Abstract

Research participation is an important component of postgraduate medical training, yet trainees often differ substantially in access to mentorship, protected research time, funding, data, infrastructure, methodological support, and collaboration. Observed scholarly output may therefore reflect not only individual characteristics but also the research-resource environment in which training occurs. We introduce Resource-Attributable Research Loss (RARL-R), a counterfactual framework for quantifying potentially remediable research-opportunity gaps in postgraduate medical education.

For trainee i, R_i denotes the observed research-resource profile, R* a prespecified feasible adequacy benchmark, and $R_i\dagger = R_i \vee R^*$ a no-downgrade remediation profile. With $\mu_i(r) = E[Y_i(r) \mid X_i]$, the primary estimand is the resource-attributable expected output gap:

$$\Delta_i^R = [\, \mu_i(R_i\dagger) - \mu_i(R_i) \,]_+$$

When the scholarly outcome has a meaningful zero and ratio-scale interpretation, this gap may additionally be expressed as equivalent research time:

$$\tau_i^R = T\, [\, 1 - \mu_i(R_i) \,/\, \mu_i(R_i\dagger) \,]_+$$

RARL-R does not estimate innate research ability, prescribe how much a trainee should have produced, or provide an adjusted ranking. Instead, it defines an individualized conditional expected counterfactual contrast between observed research resources and a feasible remediation of specified resource deficits.

We develop the conceptual and mathematical formulation of RARL-R, specify its measurement and identification requirements, and evaluate estimator recoverability using Monte Carlo simulation. Across 200 replications, a correctly specified g-computation estimator achieved mean RMSEs of 0.209, 0.132, and 0.092 equivalent research years at sample sizes of n = 200, 500, and 1000, respectively. Corresponding mean correlations between true and estimated RARL-R values were 0.938, 0.976, and 0.988. A deliberately crude cohort-reference comparator showed substantially poorer performance.

RARL-R provides an opportunity-sensitive framework for studying how modifiable features of postgraduate medical research environments may contribute to differences in expected scholarly development. The simulation demonstrates internal recoverability under a controlled, correctly

specified setting, not empirical validity in real trainees. Future evaluation requires prospective multicentre longitudinal validation, prespecified context-sensitive benchmarks, and careful assessment of confounding, positivity, measurement error, model misspecification, and interference.

Keywords: postgraduate medical education; research opportunity; scholarly development; educational equity; counterfactual inference; g-computation; research environment; equivalent research time

1. Introduction

Research and scholarly activity are increasingly regarded as important components of postgraduate medical training. Residents and other postgraduate medical trainees are commonly expected to participate in research, develop methodological competence, present scientific work, and contribute to peer-reviewed publications. Yet these expectations are pursued within markedly different research environments. Trainees with similar formal training durations may have very different access to experienced mentors, protected research time, funding, clinical data, laboratory or digital infrastructure, methodological expertise, collaboration networks, and administrative support. Consequently, observed scholarly output may reflect not only individual characteristics and effort, but also the opportunity structure within which research is undertaken.

Evidence from postgraduate medical education consistently suggests that these opportunity conditions matter. A structured review of 85 studies found that trainee publication was associated with mentor characteristics and supportive academic environments, while organised research programmes and protected research time were also associated with greater scholarly productivity [1]. In a survey of 333 medical and surgical residents, clinical time pressure, insufficient financial support, and limited access to research data were identified as important barriers to participation in research [2]. Similarly, a systematic review and meta-analysis of graduate medical education scholarly-activity initiatives reported higher publication rates following programme-level interventions, with mentoring, research curricula, and protected time among the most commonly implemented components [3]. Together, these findings indicate that scholarly development during medical training is partly contingent on modifiable features of the educational environment.

These differences may also accumulate over time. Early access to supervision, funding, protected time, and collaboration can facilitate initial research outputs, while those outputs may subsequently improve access to further opportunities. This mechanism is consistent with the broader concept of cumulative advantage in scientific careers [4–6]. For example, a funding-threshold quasi-experiment showed that early-career researchers who narrowly obtained funding subsequently accumulated substantially more funding than otherwise similar applicants who fell just below the threshold [7]. Related research in academic chemistry has also examined the relationship between research and development funding and scientific productivity [8]. Although postgraduate medical training represents a more bounded educational setting, the same underlying concern remains relevant: research opportunity and research output may become mutually reinforcing. Evaluating trainees only

on realised scholarly output may therefore obscure potentially remediable differences in the environments that helped generate that output.

Existing medical education research has made substantial progress in identifying barriers to scholarly activity and evaluating interventions intended to strengthen research participation. However, most studies address one of three questions: whether research resources are associated with scholarly outcomes, whether particular programmes improve average outcomes, or which barriers trainees report experiencing. These approaches are valuable for programme evaluation, but they do not directly define an opportunity-adjusted estimand for a particular trainee profile. In particular, they do not answer the following counterfactual question: conditional on a trainee's measured pre-resource characteristics, how much would expected scholarly output differ if specified and modifiable research-resource deficits were improved to a feasible adequacy level? Broader higher-education scholarship has also examined changing conceptions of inequality and unequal access to elite postgraduate education, providing context for the structural dimension of these opportunity differences [9,10].

This distinction is important because commonly used comparison groups cannot answer that question. Comparing a trainee with the cohort mean, a high-performing peer group, or trainees from a better-resourced institution combines differences in research opportunity with differences in prior preparation, motivation, specialty, stage of training, career intention, and other baseline characteristics. The resulting gap therefore cannot be interpreted specifically as a resource-attributable opportunity gap. A more appropriate approach is to compare two conditional expected outcomes for the same trainee profile: the expected outcome under the observed research-resource configuration and the expected outcome under a clearly defined resource-remediation intervention. Such a comparison does not recover an unobservable deterministic outcome for an individual, nor does it infer innate research ability. Rather, it defines an individualized conditional expected counterfactual contrast under explicit assumptions about the intervention, measured covariates, and data-generating process.

We therefore introduce Resource-Attributable Research Loss (RARL-R), a counterfactual framework for quantifying potentially remediable research-opportunity deficits in postgraduate medical training. Let R_i denote the observed research-resource profile for trainee i, and let R* denote a prespecified feasible adequacy benchmark. RARL-R constructs a no-downgrade remediation profile in which resource domains below the benchmark are raised to the adequacy level while resources already exceeding the benchmark are preserved. The framework then compares the trainee-specific conditional expected scholarly outcome under the observed resource profile with the corresponding expected outcome under the remediated profile. The difference defines a resource-attributable expected output gap. Under additional assumptions, this gap may also be expressed as an equivalent amount of research time under the benchmark resource condition to facilitate interpretation.

RARL-R is intended primarily as a framework for understanding and evaluating educational environments rather than for ranking individual trainees. It does not specify what a trainee should have produced, assume that individuals would achieve identical outcomes under identical resources, or replace qualitative assessment of research quality. Instead, it asks a narrower and potentially actionable question: to what extent might expected scholarly development differ if identified deficits

in mentorship, protected time, funding, data access, methodological support, infrastructure, or collaboration were feasibly remediated? In this sense, RARL-R complements existing research on barriers and scholarly-activity interventions by translating research opportunity into an explicitly defined counterfactual target that can, in principle, be evaluated using longitudinal or quasi-experimental data.

Although the present paper operationalises RARL-R in postgraduate medical education, the conceptual framework is not inherently restricted to medical training. Its underlying logic may be applicable to other research and higher-education settings in which resource conditions are measurable, opportunity deficits are potentially remediable, and meaningful outcome contrasts can be defined. Postgraduate medical training provides an initial application domain in which research expectations are explicit and variation in mentorship, protected time, funding, data access, and methodological support can be examined. Extension to other settings, however, requires context-specific definitions of resource domains, feasible adequacy benchmarks, and scholarly outcomes rather than assuming a universal resource profile. Accordingly, the present study should be understood as an initial operationalisation of a broader opportunity-sensitive research framework, rather than a claim that RARL-R is exclusively a medical education measure.

The aim of this study is to develop the conceptual, mathematical, and methodological foundations of RARL-R for postgraduate medical education. Specifically, we: (1) define research opportunity as a multidimensional and potentially modifiable educational exposure; (2) specify a feasible no-downgrade resource-remediation intervention and the corresponding counterfactual estimand; (3) state the identification assumptions required to interpret RARL-R as resource-attributable rather than merely resource-associated; (4) evaluate estimator recoverability using Monte Carlo simulation; and (5) outline an empirical validation strategy for future multicentre studies. By shifting attention from observed scholarly output alone to the conditions under which that output becomes attainable, RARL-R aims to provide medical educators and programme leaders with a quantitative framework for examining potentially remediable inequalities in research training.

2. Conceptual Framework

RARL-R is based on the premise that scholarly development during postgraduate medical training emerges from the interaction between trainee characteristics and the research environment in which training occurs. Observed scholarly output is therefore treated as an achieved outcome rather than a direct measure of intrinsic research capability. The framework focuses specifically on research opportunities that are plausibly modifiable through educational, programme-level, or institutional action.

The purpose of RARL-R is not to represent every determinant of research development. Motivation, career preference, prior training, specialty, personal circumstances, and other individual factors may all influence scholarly outcomes. Instead, the framework isolates a narrower set of resource conditions for which a meaningful intervention can be specified. Its central question is whether expected scholarly development would differ if identified and remediable deficits in those conditions were improved to a feasible level.

2.1 Research opportunity as a modifiable educational exposure

We define a trainee's research opportunity as the multidimensional set of resources and institutional conditions that enable, constrain, or shape participation in scholarly activity during postgraduate medical training. Candidate domains include mentorship, protected research time, financial support, access to clinical data and research infrastructure, methodological and statistical support, and opportunities for collaboration.

This definition deliberately distinguishes research opportunity from institutional reputation or prestige. Trainees within the same hospital, university, department, or residency programme may experience substantially different research environments. Conversely, a trainee in a less research-intensive institution may have strong access to particular resources. The relevant exposure is therefore the resource profile actually available to the trainee rather than a global label assigned to the institution.

Research opportunity also has multiple levels of availability. A resource may be formally provided without being practically accessible or effectively usable. Protected research time, for example, may appear in a programme schedule but be repeatedly interrupted by clinical responsibilities. Statistical support may formally exist but involve waiting periods that make it unavailable within the relevant research timeline. Similarly, assignment to an academically productive supervisor does not necessarily imply timely project-level mentorship.

For this reason, empirical applications of RARL-R should distinguish where possible between nominal provision, realised access, and effective usability. Resource domains should be measured using transparent and reproducible indicators, with objective administrative measures preferred when available and validated trainee-reported measures used when they capture aspects of access or quality that cannot otherwise be observed. Related scholarship on doctoral education and early-career researcher development, although not specific to medical residency, has also emphasized the role of supervision and researcher-development trajectories [11,12].

Let the observed research-resource profile for trainee i be represented as

$$R_i = (R_{i1}, R_{i2}, ..., R_i \quad) \tag{1}$$

where each component represents a prespecified and potentially modifiable domain of research opportunity.

RARL-R does not assume that an identical resource profile would generate identical scholarly outcomes for all trainees. The conversion of research opportunity into scholarly development may vary according to pre-resource characteristics such as prior research experience, specialty, training stage, methodological preparation, and career goals. These characteristics are represented by X_i and are held conceptually fixed when comparing the trainee's expected outcomes under alternative resource conditions.

Accordingly, RARL-R is not a comparison between different trainees. It compares conditional expected outcomes for the same trainee profile under two explicitly defined research-resource states.

2.2 Feasible resource remediation

RARL-R requires an intervention that is sufficiently explicit to support counterfactual interpretation. The relevant intervention is not universal resource equalisation and does not assume that all trainees should receive identical research environments. Instead, the framework defines a feasible adequacy benchmark for each resource domain.

Let

$$R^* = (R_1^*, R_2^*, ..., R\quad^*) \tag{2}$$

denote a prespecified resource-adequacy benchmark.

The benchmark represents a defensible minimum or adequate level of research opportunity rather than an ideal, maximal, or elite resource environment. In empirical studies, R* may be informed by accreditation standards, programme requirements, expert consensus, empirical distributions among comparable programmes, or other externally justified criteria. The benchmark should be specified before outcome modelling wherever possible and should be subjected to sensitivity analysis.

For each trainee, the resource-remediated profile is defined component-wise as

$$R_i\dagger = R_i \vee R^* \tag{3}$$

where $\vee$ denotes the component-wise maximum.

Thus,

$$R_i\quad\dagger = \max(R_i\quad, R\quad^*) \tag{4}$$

This no-downgrade rule has an important conceptual implication. Resource domains below the adequacy benchmark are raised to the benchmark, whereas domains already above it are left unchanged. The intervention therefore represents remediation of specified opportunity deficits rather than redistribution toward uniformity.

For example, a trainee with adequate methodological support but insufficient protected research time would receive a counterfactual improvement only in protected time. RARL-R would not reduce methodological support simply because it exceeds the benchmark.

The choice of R* is therefore substantive rather than purely statistical. A benchmark that is unrealistic, excessively high, or poorly aligned with the training context would generate a counterfactual contrast of limited educational relevance. Empirical applications should consequently report the rationale for the selected benchmark and examine whether conclusions are robust across alternative plausible adequacy thresholds.

2.3 Counterfactual estimand

Let $Y_i(r)$ denote the potential scholarly outcome for trainee i under research-resource profile r. Rather than attempting to recover a deterministic individual outcome, RARL-R focuses on the conditional expected potential outcome [13–15]

$$\mu_i(r) = E[Y_i(r) \mid X_i] \tag{5}$$

where X_i represents measured pre-resource characteristics.

The framework compares two conditional expected outcomes: the expected scholarly outcome under the observed resource profile, $\mu_i(R_i)$, and the expected scholarly outcome under the feasible resource-remediation profile, $\mu_i(R_i\dagger)$.

The resource-attributable expected output gap is defined as

$$\Delta_i^R = [\mu_i(R_i\dagger) - \mu_i(R_i)]_+ \tag{6}$$

where the positive-part function is defined as

$$[z]_+ = \max(z, 0) \tag{7}$$

The positive-part operator prevents an estimated decrease in output following resource remediation from being labelled a resource-attributable "loss". Such findings should instead prompt examination of model specification, measurement, non-monotonic resource effects, or the appropriateness of the proposed remediation intervention.

The estimand should therefore be interpreted as an individualized conditional expected counterfactual contrast, not as the unobserved true amount of research that a particular trainee would certainly have produced under different circumstances. Its interpretation depends on the validity of the resource intervention, the adequacy of measured baseline characteristics, and the causal identification assumptions described subsequently.

RARL-R can additionally translate the expected output gap into equivalent research time. Let T denote the observation or training horizon. Under the simplifying assumption that expected productivity under the remediated condition is constant over this period, the equivalent research-time measure is

$$\tau_i^R = T\,[1 - \mu_i(R_i) / \mu_i(R_i\dagger)]_+ \tag{8}$$

The equivalent-time representation is intended as an interpretive transformation rather than the primary causal estimand. A value of 0.5 years does not imply that a trainee literally lost six months of calendar time. Rather, under the specified model and benchmark, the expected output gap associated with remediable resource deficits corresponds to the amount of production that would be expected during 0.5 years under the benchmark resource condition.

The primary quantity of interest therefore remains the resource-attributable expected output gap, while equivalent research time provides a secondary scale for communicating its magnitude.

This formulation distinguishes RARL-R from conventional trainee ranking or productivity scoring. The framework does not estimate innate research potential, prescribe how much a trainee ought to produce, or assume that equal resources would eliminate individual differences. It defines a specific intervention-based contrast: how much would expected scholarly development differ,

conditional on measured baseline characteristics, if specified research-opportunity deficits were feasibly remediated?

2.4 Mathematical properties and interpretive safeguards

The mathematical properties of RARL-R are summarized in Table 1. These properties are intended to ensure that the estimand behaves consistently under the proposed resource-remediation framework. They include non-negativity, zero-loss consistency, boundedness under productive remediation, positive scale invariance, no-downgrade benchmark consistency, and monotonicity when the specified resource intervention weakly increases expected scholarly output.

These properties describe the internal behaviour of the estimand; they do not establish causal validity. A mathematically well-behaved RARL-R value may still be misleading if the resource intervention is poorly defined, relevant confounders are unmeasured, positivity is violated, resource or outcome measures are unreliable, or interference between trainees is ignored.

The equivalent-time transformation also requires additional interpretive caution. It is meaningful only when the scholarly outcome has a non-negative ratio scale with a meaningful zero. When this requirement is not satisfied, the resource-attributable expected output gap should remain the primary estimand and should not be converted into equivalent research time.

Figure 1 summarizes the conceptual architecture of the framework: measured pre-resource characteristics and observed research opportunities inform expected scholarly output, while the adequacy benchmark defines a no-downgrade remediated resource profile. Expected outcomes under the observed and remediated profiles are then contrasted to obtain the resource-attributable expected output gap and, where appropriate, its equivalent-time representation.

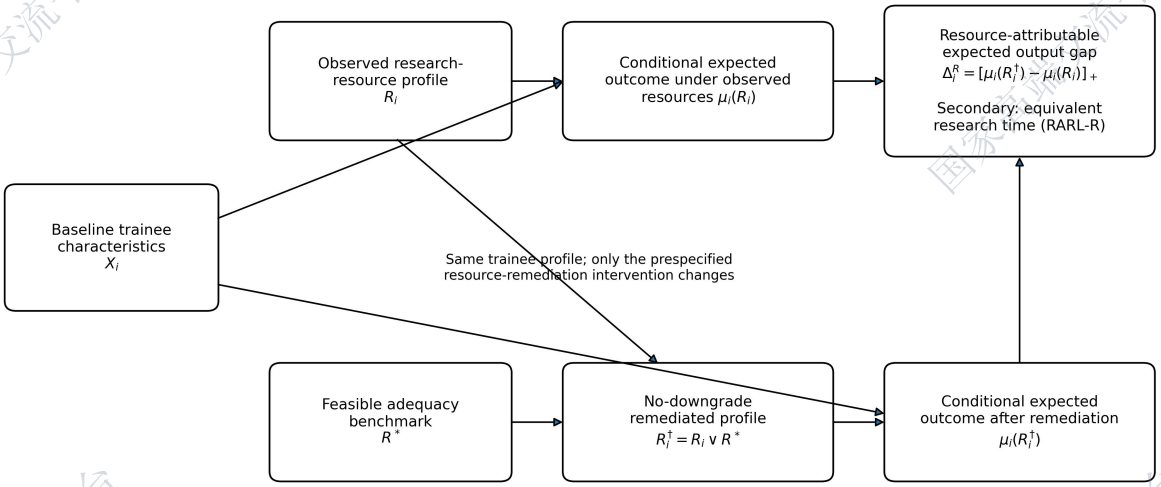


Figure 1. Counterfactual architecture of RARL-R in postgraduate medical research training.

Alt text: Baseline trainee characteristics and observed research resources contribute to expected scholarly output. A feasible adequacy benchmark is combined with the observed resource profile to construct a no-downgrade remediated profile. Conditional expected outcomes under the observed and remediated profiles are compared to estimate the resource-attributable expected output gap, which may subsequently be translated into equivalent research time.

3. Measurement Architecture for Empirical RARL-R Studies

Empirical implementation of RARL-R requires more than fitting an outcome model. The resource intervention, scholarly outcome, baseline covariates, and adequacy benchmark must each be defined before estimation in a way that corresponds to a meaningful educational question. The validity of RARL-R therefore depends fundamentally on measurement design.

3.1 Research-resource domains

The resource vector should contain domains that are conceptually relevant to postgraduate medical research development and sufficiently modifiable to support an educational or institutional intervention. Candidate domains include mentorship, protected research time, direct research funding, access to clinical data and research infrastructure, methodological and statistical support, and opportunities for collaboration. Illustrative indicators are summarized in Table 2.

Resource measurement should distinguish, where possible, between three related but different constructs: nominal provision, realised access, and effective usability. For example, a residency programme may formally provide protected research time, but repeated interruption by clinical duties may substantially reduce the time that is actually available. Similarly, statistical consultation may exist institutionally but remain functionally inaccessible because of long waiting times or restrictive eligibility criteria.

Accordingly, empirical studies should not infer individual opportunity solely from institutional availability. Administrative data should be used where feasible—for example, scheduled protected hours, programme-level funding, waiting times for methodological support, or access permissions for clinical databases—and should be complemented by validated trainee-reported measures when important features such as supervisory responsiveness or practical accessibility cannot otherwise be captured.

Resource definitions should be specified before outcome analysis. Investigators should also state whether each variable represents availability, access, intensity, or quality, because these constructs are not interchangeable.

3.2 Scholarly outcomes

The primary outcome should represent scholarly development during a clearly defined observation window and should be sufficiently transparent to permit educational interpretation.

Potential outcomes include numbers of peer-reviewed research articles, completion of predefined research projects, registered protocols, datasets or other research products, competitive research awards, conference presentations, and other programme-relevant scholarly achievements. No single outcome is suitable for every training setting, and the primary measure should therefore reflect the objectives of the programme being evaluated.

Use of journal-level metrics as proxies for individual research quality should be avoided. In particular, raw Journal Impact Factor should not be treated as an individual-level scholarly-performance measure. Where multiple forms of contribution are combined, weighting rules should be prespecified and justified rather than chosen after observing the results [16–20].

A pragmatic first empirical validation of RARL-R could therefore use either a simple count outcome or a transparent prespecified Research Contribution Score, accompanied by sensitivity analyses using separate domain-specific outcomes. Because composite weighting schemes inevitably contain normative choices, results should also be reported using simpler outcomes such as publication counts or completed research projects.

The equivalent research-time transformation imposes a stronger measurement requirement than the expected output-gap estimand itself. It should be applied only to non-negative outcomes with a meaningful zero and ratio-scale interpretation. If an ordinal score, latent construct, or other non-ratio outcome is used, investigators should report the expected output gap without converting it into equivalent research years.

3.3 Pre-resource covariates and temporal ordering

Let X_i contain characteristics measured before, or clearly antecedent to, the resource exposure being evaluated. Candidate variables may include prior research experience, previous publications, methodological training, specialty, stage of postgraduate training, relevant academic qualifications, and other factors that plausibly influence both subsequent access to research resources and scholarly outcomes.

Temporal ordering is critical. Variables that arise after resource allocation, or that are themselves affected by the resource environment, should not automatically be adjusted for as baseline confounders. For example, research self-efficacy measured after sustained mentorship may partly mediate the effect of mentorship and therefore should not be treated as an ordinary pre-exposure covariate without an explicit causal rationale.

The appropriate covariate set should be determined by the assumed causal structure rather than by automated variable-selection procedures or associations with the outcome alone. Directed acyclic graphs may be useful for documenting these assumptions before analysis [14,15].

When institutional or programme characteristics influence both resource allocation and scholarly outcomes, these contextual variables should also be considered in the design and modelling strategy. Multicentre studies should distinguish individual-level resources from programme- and institution-level determinants rather than combining them indiscriminately.

3.4 Specification of the adequacy benchmark

The adequacy benchmark R^* is a substantive component of the intervention and should not be treated as a tuning parameter selected to optimise model performance or maximise the estimated RARL-R value.

Benchmarks may be defined using external programme standards, accreditation requirements, policy entitlements, expert consensus, or empirically defensible thresholds derived from comparable training programmes. If empirical distributions are used, benchmark selection should ideally be performed without reference to scholarly outcomes.

Benchmarks may differ across specialties or stages of training when the educational meaning of a resource differs materially. For example, appropriate protected research time for a laboratory-based doctoral trainee may not be directly transferable to a clinically intensive residency programme.

Because no single adequacy threshold is universally correct, empirical RARL-R studies should prespecify a primary benchmark and evaluate sensitivity to alternative plausible values. For continuous resource scales, analyses might compare lower, intermediate, and higher feasible adequacy thresholds. Conclusions that depend strongly on a narrow benchmark choice should be interpreted cautiously.

The benchmark should therefore answer a practical educational question: what level of this resource could realistically be regarded as adequate and achievable for trainees in this context?

4. Identification and Estimation

RARL-R distinguishes the definition of the target estimand from the statistical procedure used to estimate it. The counterfactual contrast is defined conceptually before any regression model is fitted. Statistical estimation can recover a resource-attributable interpretation only when additional identification assumptions are credible.

4.1 Conditions for causal interpretation

The conditions required to interpret RARL-R as resource-attributable rather than merely resource-associated are summarized in Table 3.

Consistency requires that the observed outcome under the trainee's observed resource profile corresponds to the potential outcome defined under that same resource profile. This assumption depends on sufficiently precise specification of the resource intervention. Broad constructs such as "good mentorship" are therefore inadequate unless translated into interpretable and measurable intervention components.

Conditional exchangeability requires that, after conditioning on the selected pre-resource covariates, there are no important unmeasured common causes of resource exposure and potential scholarly outcomes. This assumption is particularly demanding in observational educational research because motivation, prior supervisor relationships, programme selection, and institutional culture may influence both resource access and subsequent research performance [13–15].

Positivity requires empirical support for the resource contrasts being estimated among trainees with relevant covariate patterns. If particular trainee profiles are observed only in poorly resourced or only in highly resourced environments, the corresponding counterfactual comparison relies heavily on extrapolation and may not be empirically identifiable.

Temporal ordering requires that confounders precede the relevant resource exposure and that variables caused by the resource environment are not naively adjusted as though they were baseline characteristics.

Measurement validity requires sufficiently reliable and comparable measurement of both research resources and scholarly outcomes. Systematic misclassification of resource access may substantially distort the estimated remediation contrast.

Interference also deserves particular attention. Trainees may share supervisors, statistical services, databases, laboratory facilities, and collaborative networks. Improving resources for one trainee may therefore influence the opportunities or outcomes of others. Where such spillovers are plausible, conventional no-interference assumptions may be inappropriate and cluster- or network-aware designs should be considered.

These assumptions cannot be established through statistical fit alone. When the design does not support a credible causal interpretation—for example, in a simple cross-sectional survey—the resulting quantity should be described as an opportunity-associated rather than resource-attributable expected output gap [14,15].

4.2 G-computation estimation

Let

$$m(x, r) = E(Y \mid X = x, R = r) \tag{9}$$

denote the conditional outcome regression.

After estimating m(x,r), the expected outcome for trainee i under the observed resource profile is

$$\hat{\mu}_i(R_i) = \hat{m}(X_i, R_i) \tag{10}$$

whereas the expected outcome under the no-downgrade remediation profile is

$$\hat{\mu}_i(R_i\dagger) = \hat{m}(X_i, R_i\dagger) \tag{11}$$

The estimated resource-attributable expected output gap is therefore

$$\hat{\Delta}_i^R = [\hat{\mu}_i(R_i\dagger) - \hat{\mu}_i(R_i)]_+ \tag{12}$$

For outcomes satisfying the requirements for the equivalent-time transformation,

$$\hat{\tau}_i^R = T\,[1 - \hat{\mu}_i(R_i) / \hat{\mu}_i(R_i\dagger)]_+ \tag{13}$$

These equations describe g-computation as an estimator of the previously defined counterfactual quantities. They do not redefine the estimand itself [15].

This distinction is fundamental. High predictive accuracy does not establish causal identification, and the use of sophisticated machine-learning algorithms cannot compensate for unmeasured confounding, poor positivity, post-treatment adjustment, invalid measurement, or an ill-defined remediation intervention.

4.3 Model specification and flexible estimation

The appropriate outcome model should reflect the distribution and structure of the prespecified scholarly outcome. Count outcomes may require Poisson, negative-binomial, or other appropriate models, whereas positive continuous outcomes may require alternative generalized linear models. Model choice should be justified by the outcome rather than selected solely according to which specification produces the largest resource-attributable contrast.

In an initial empirical validation, a prespecified and interpretable regression model may provide the clearest test of the framework. More flexible approaches can subsequently be used when nonlinearities and interactions between resource domains are plausible.

Machine-learning methods may be incorporated into g-computation, preferably with procedures such as sample splitting or cross-fitting where appropriate, but prediction and causal estimation should remain conceptually distinct. Flexible modelling may reduce functional-form misspecification; it does not resolve violations of the underlying identification assumptions.

Doubly robust estimators may also be considered when the resource intervention can be operationalised in a form for which both outcome and exposure models can be meaningfully specified. Their use becomes more complicated when R consists of multiple continuous and interacting resource domains and should therefore be justified rather than adopted automatically.

4.4 Positivity, sensitivity analyses, and uncertainty

Empirical applications should examine whether the remediated resource profiles remain within regions supported by the observed data. Diagnostics should therefore evaluate overlap and identify trainees for whom estimation depends heavily on extrapolation. Where severe positivity problems occur, investigators may need to redefine the benchmark, restrict the target population, or reconsider the intervention rather than rely on increasingly complex prediction models [15].

Sensitivity analyses should address at least four sources of uncertainty where feasible: alternative adequacy benchmarks, alternative outcome definitions, plausible resource-measurement error, and alternative outcome-model specifications. In observational studies, additional analyses addressing potential unmeasured confounding should be considered.

Uncertainty intervals should reflect the complete estimation procedure. Nonparametric bootstrap procedures may be used when appropriate, with resampling performed at the programme or institutional level when observations are clustered. If R^* is empirically estimated rather than externally fixed, benchmark estimation should be repeated within each bootstrap sample rather than treated as known.

Missing data procedures should likewise be incorporated into the full estimation workflow. When multiple imputation is used, resource-remediation construction and counterfactual estimation should be repeated within each imputed dataset before uncertainty estimates are combined.

The objective of these procedures is not to produce a single apparently precise RARL-R value, but to determine whether the substantive conclusion—that specified and remediable differences in research opportunity are associated with meaningful differences in expected scholarly development—is robust to plausible analytical choices.

5. Interpretive Implications and Safeguards for Educational Use

The counterfactual formulation of RARL-R has implications for how estimates should be interpreted in postgraduate medical education. These implications are not additional causal assumptions or empirical findings. Rather, they define the intended scope of the framework and provide safeguards against uses that would be inconsistent with its estimand.

5.1 RARL-R quantifies remediable opportunity gaps, not trainee deficiency

A positive RARL-R value should not be interpreted as evidence of low ability, insufficient motivation, or poor individual performance. The framework compares conditional expected scholarly outcomes under two resource conditions while holding measured pre-resource characteristics fixed. The resulting contrast therefore concerns the specified research environment and remediation intervention rather than an intrinsic property of the trainee.

Trainees with similar observed scholarly output may have different estimated resource-attributable gaps because they developed within different research environments. Conversely, trainees with different observed outputs may have similar RARL-R values if the estimated contribution of remediable resource deficits is comparable.

RARL-R should therefore not be used as a substitute for assessment of research quality, competence, motivation, or academic merit. In particular, it is not designed for trainee ranking, selection, promotion, credentialing, or punitive performance management. Its primary educational use is to identify potentially modifiable features of research training environments [17–20].

5.2 Resource adequacy is not equivalent to resource equality

RARL-R does not assume that every trainee should receive identical research resources. If all components of a trainee's observed resource profile already meet the prespecified adequacy benchmark, then

$$R_i\dagger = R_i \tag{14}$$

and the corresponding resource-attributable gap is zero, even if other trainees receive substantially greater resources.

The framework therefore distinguishes between unequal resources and inadequate resources. Only prespecified domains falling below a defensible adequacy threshold are modified by the no-downgrade intervention.

This distinction is important for educational interpretation. RARL-R is intended to evaluate potentially remediable shortfalls in research opportunity, not to define every difference between programmes or trainees as inequitable. The normative meaning of “adequacy” must therefore be justified in the educational context rather than inferred solely from relative ranking within a cohort.

5.3 Programme averages may conceal the distribution of research opportunity

Aggregate measures of scholarly productivity may obscure substantial within-programme heterogeneity in research opportunity. Two residency or postgraduate programmes could achieve similar average publication rates while exposing their trainees to very different distributions of mentorship, protected time, methodological support, data access, or collaboration.

For example, one programme may provide broadly adequate resources across most trainees, whereas another may generate a similar mean output through very strong support for a small subgroup alongside substantial deficits for others. Programme-level mean scholarly output alone cannot distinguish these configurations.

For this reason, empirical applications of RARL-R should not reduce results to a single institutional score. Where sample size permits, programme evaluation should describe the distribution of estimated resource-attributable gaps and the prevalence of deficits across resource domains. Relevant summaries may include the median and interquartile range of RARL-R, the proportion of trainees below each adequacy benchmark, and domain-specific patterns of remediation need.

Such summaries are more consistent with the purpose of RARL-R than institutional league tables, because the primary objective is to identify potentially remediable features of the educational environment.

5.4 The conversion of resources into scholarly outcomes may be heterogeneous

RARL-R does not require an assumption that the same resource intervention produces the same effect for every trainee. The relationship between research opportunity and scholarly development may vary according to prior research experience, specialty, stage of training, methodological preparation, career goals, and other baseline characteristics.

Resource domains may also interact. Additional protected research time, for example, may have limited value when trainees lack access to mentorship or data, whereas the same increase in protected time may have a larger effect in an environment where complementary resources are already available.

The individualized conditional structure of RARL-R permits such heterogeneity in principle. Empirical models should therefore consider scientifically plausible nonlinearities and interactions when supported by study design and sample size. At the same time, subgroup effects should not be generated through unrestricted exploratory modelling, particularly in small educational datasets.

Average resource effects should consequently not be interpreted as universal response functions. The educational question is not whether a particular resource is equally productive for everyone, but whether a defined remediation changes expected scholarly outcomes for trainees with particular measured baseline characteristics.

5.5 RARL-R estimates must remain empirically falsifiable

RARL-R should generate claims that can be challenged by data. A resource domain should not automatically be regarded as causally important simply because it is conceptually desirable or educationally valued.

If longitudinal, experimental, or credible quasi-experimental evidence indicates that improving a specified resource produces little or no change in expected scholarly outcomes, then the estimated resource-attributable gap associated with that domain should correspondingly be small. The framework should not preserve a large estimated loss through post hoc claims about unmeasured or unrealised potential.

Similarly, a remediation contrast that requires substantial extrapolation beyond observed resource combinations should not be presented with the same confidence as one supported by adequate empirical overlap. Severe positivity violations, unstable estimates, sensitivity to benchmark specification, or major dependence on model assumptions should reduce the strength of substantive conclusions.

These constraints are central to the intended role of RARL-R. The framework is not designed to demonstrate that resource inequality must explain observed differences in scholarly output. It provides a formal way to test how much expected scholarly development may be associated with specified and remediable features of the research environment under explicit assumptions.

5.6 Educational interpretation should prioritise remediation over ranking

The most direct use of RARL-R is therefore diagnostic rather than classificatory. At programme level, the framework may help identify whether expected scholarly opportunity gaps are concentrated in mentorship, protected research time, methodological support, data infrastructure, funding, collaboration, or combinations of these domains.

This distinction matters because different opportunity deficits imply different educational responses. A programme in which protected time is the principal constraint may require changes in clinical scheduling, whereas a programme with adequate time but limited methodological support

may benefit more from biostatistical consultation, research-methods training, or shared analytic infrastructure.

RARL-R is therefore most informative when estimates are linked to potentially actionable resource domains. Its educational value lies not in producing a new score by which trainees or institutions can be ranked, but in making explicit which modifiable conditions may contribute to differences in expected scholarly development and which interventions merit further evaluation.

6. Monte Carlo Simulation Study

The purpose of the simulation study was to examine whether the proposed RARL-R estimand can be recovered under a controlled data-generating process in which the true counterfactual quantities are known. The simulation was designed as a proof-of-concept evaluation of estimator recoverability rather than as a model of any specific postgraduate medical training system.

Accordingly, the results should not be interpreted as evidence that real trainees experience a particular magnitude of resource-attributable research loss. They address a narrower methodological question: when the resource-remediation intervention, outcome-generating mechanism, and relevant baseline covariates are known, can an estimator based on g-computation recover the corresponding RARL-R values?

6.1 Data-generating mechanism

Each simulated trainee had two pre-resource baseline characteristics and four correlated research-resource domains scaled to the interval (0, 1). The four resource domains were intended to represent generic components of the postgraduate medical research environment rather than literal measures of specific programmes.

Resource values were generated from a shared latent factor, resource-specific random variation, and observed baseline characteristics, followed by logistic transformation to constrain all values to (0, 1). This structure induced correlation between resource domains while allowing baseline trainee characteristics to influence resource access.

Expected scholarly output was generated from a positive log-link function incorporating baseline characteristics, the four resource domains, and one prespecified interaction between resource domains. Observed outcomes were then drawn from a Gamma distribution with shape parameter 5 and conditional mean equal to the generated expected output.

The primary feasible adequacy benchmark was specified as

$$R^* = (0.75, 0.75, 0.75, 0.75) \tag{15}$$

and the observation horizon was set to

$$T = 3 \text{ years} \tag{16}$$

For each simulated trainee, the true remediated resource profile was constructed using the no-downgrade rule

$$R_i\dagger = R_i \vee R^* \tag{17}$$

Because the complete data-generating mechanism was known, the true conditional expected outcomes under both the observed and remediated resource profiles could be calculated directly. These quantities were used to derive the true individual resource-attributable expected output gap and true RARL-R value.

The complete data-generating parameters, fixed random seed, and executable simulation code are provided in Supplementary Material S1 and Supplementary Code S1.

6.2 Estimation strategy and comparator

A Gamma generalized linear model with log link was fitted within each simulated dataset. In the primary proof-of-concept scenario, the fitted model matched the functional form of the data-generating mechanism.

For each trainee, g-computation was used to generate two predicted conditional expected outcomes:

$$\hat{\mu}_i(R_i) \tag{18}$$

under the observed resource profile, and

$$\hat{\mu}_i(R_i\dagger) \tag{19}$$

under the no-downgrade remediated profile.

These predictions were used to estimate the resource-attributable expected output gap and equivalent research-time measure according to the definitions introduced in Section 2.

A deliberately crude cohort-reference approach was included as a comparator. This method used the cohort-level mean predicted scholarly output as the reference value rather than constructing an individualized counterfactual under the resource-remediation intervention.

The comparator was not intended to represent an optimal alternative estimator. Its purpose was to illustrate the conceptual limitation of replacing an individualized resource counterfactual with a group-level reference.

6.3 Monte Carlo design and performance measures

Simulations were conducted at sample sizes of

$$n = 200, 500, 1000 \tag{20}$$

with 200 Monte Carlo replications at each sample size.

Estimator performance was assessed primarily using root mean squared error (RMSE) between true and estimated individual RARL-R values and Pearson correlation between true and estimated RARL-R across trainees.

RMSE was used to quantify the magnitude of individual-level estimation error, whereas correlation assessed the extent to which estimated values preserved variation in the true underlying RARL-R distribution.

These metrics were selected to evaluate internal recoverability of the individual conditional estimand. They should not be interpreted as evidence of calibration or transportability to real postgraduate medical education populations.

6.4 Simulation results

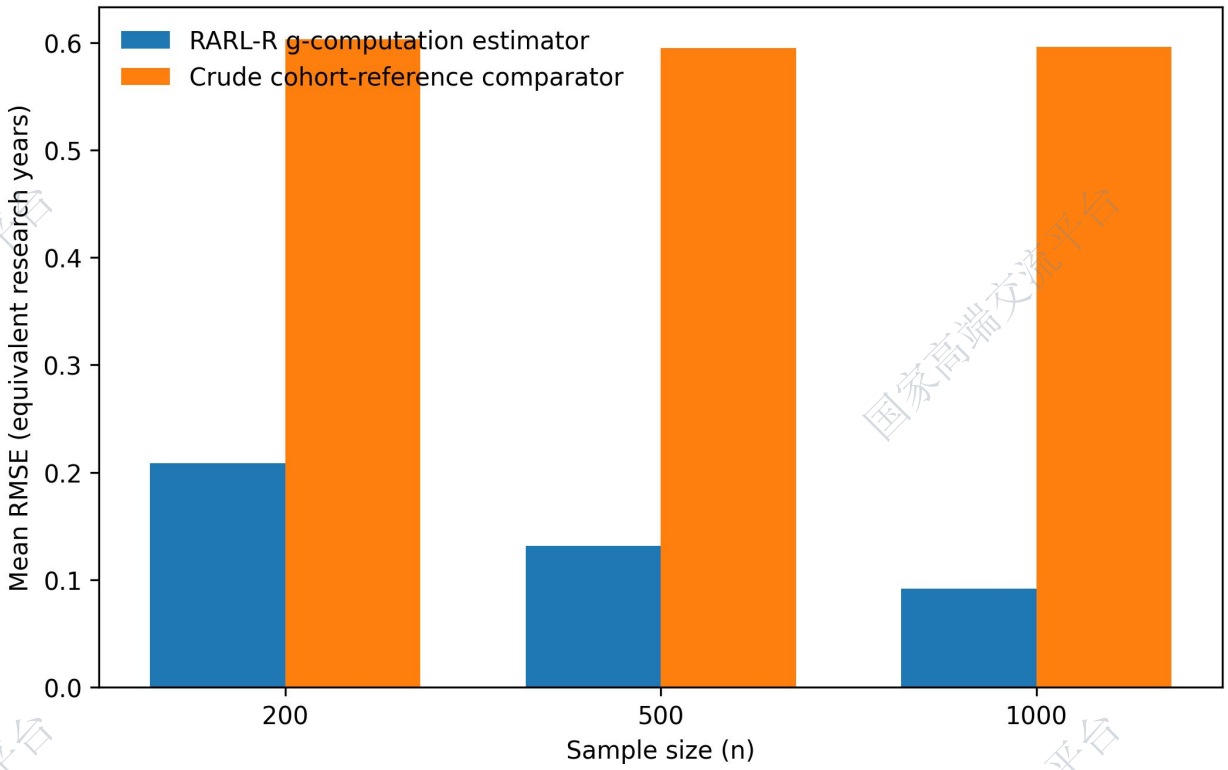
Table 4 summarizes estimator performance across the 200 replications at each sample size.

Under the correctly specified proof-of-concept scenario, recovery improved as sample size increased. Mean RMSE for the g-computation estimator decreased from 0.209 years at n = 200 to 0.132 years at n = 500 and 0.092 years at n = 1000.

The mean correlation between true and estimated individual RARL-R values increased from 0.938 at n = 200 to 0.976 at n = 500 and 0.988 at n = 1000.

By contrast, the crude cohort-reference approach showed substantially larger error. Mean RMSE remained approximately 0.60 years across sample sizes, while correlations with the true individual RARL-R values remained approximately 0.63–0.64.

Figure 2 illustrates the difference in RMSE between the counterfactual estimator and the crude cohort-reference comparator across sample sizes.



国家高端交流平台

Figure 2. Monte Carlo RMSE of the RARL-R g-computation estimator and crude cohort-reference comparator.

Alt text: Bar chart comparing mean RMSE across sample sizes 200, 500, and 1000. RMSE for the RARL-R g-computation estimator decreases as sample size increases and is consistently lower than that of the crude cohort-reference comparator.

Figure 3 provides an illustrative replication at n = 500, showing the relationship between true and estimated RARL-R values.

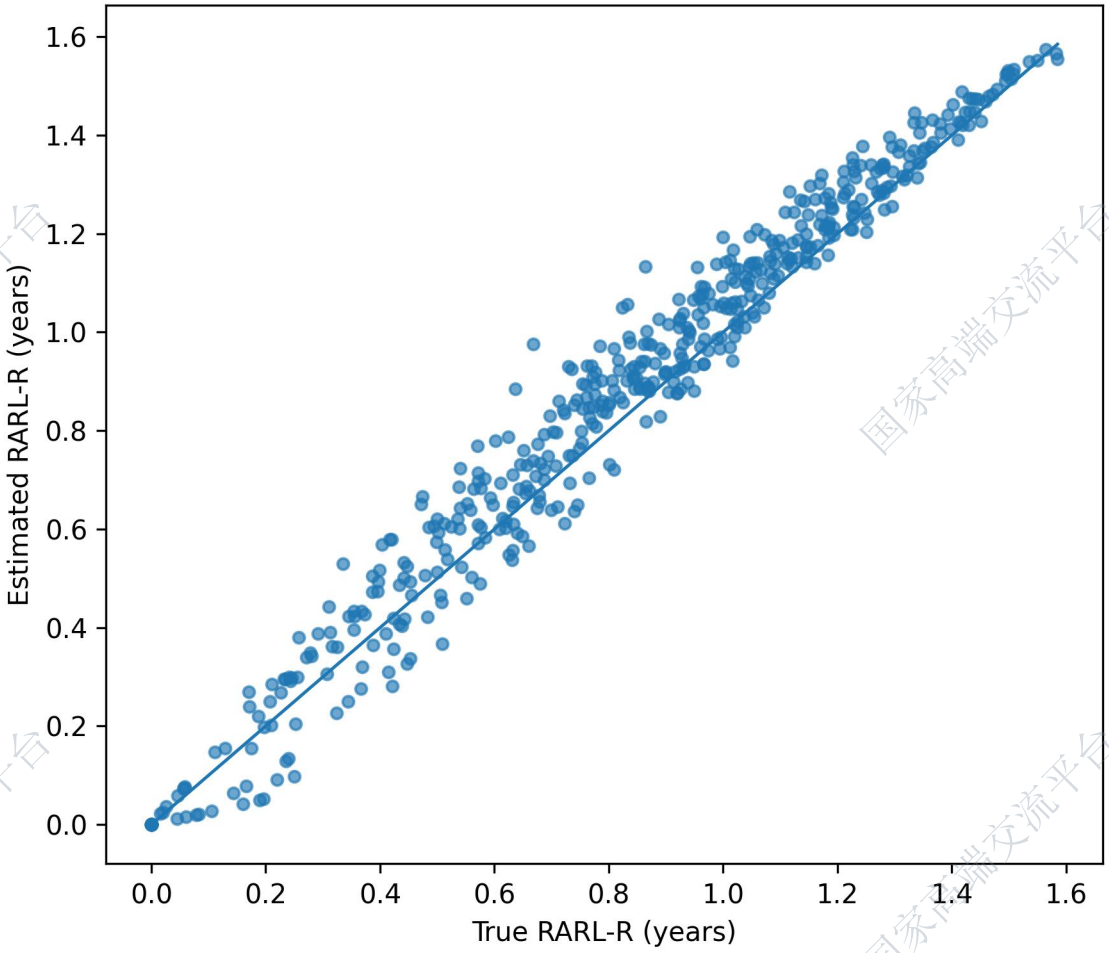


Figure 3. Example Monte Carlo replication at n = 500: true versus estimated RARL-R.

Alt text: Scatter plot comparing true and estimated RARL-R values for 500 simulated trainees. Most observations lie near the 45-degree identity line under the correctly specified proof-of-concept scenario.

6.5 Interpretation of the simulation

The simulation demonstrates that the proposed estimand is recoverable under a controlled setting in which the intervention is well defined, the relevant baseline covariates are observed, positivity is adequate, and the outcome model is correctly specified.

This finding is deliberately limited. It does not demonstrate that RARL-R is valid in real postgraduate medical education data, nor does it establish that the assumed resource-outcome relationships correspond to those operating in actual training programmes.

In particular, the primary simulation represents a favourable identification and modelling scenario. Real-world applications may involve nonlinear resource effects, model misspecification, resource-measurement error, limited overlap, unmeasured confounding, clustering, and interference between trainees.

The present Monte Carlo study should therefore be interpreted as a test of computational and statistical recoverability rather than empirical validation of the framework.

Future simulation work should explicitly stress-test RARL-R under model misspecification, weaker positivity, measurement error in research-resource variables, alternative adequacy benchmarks, and more complex interactions between resource domains. These scenarios are particularly important because the usefulness of RARL-R depends not only on whether the estimator performs under ideal conditions, but also on how rapidly performance deteriorates when assumptions resemble those encountered in real educational data.

7. Empirical Validation Blueprint for Postgraduate Medical Education

Simulation alone cannot establish whether RARL-R represents meaningful differences in research opportunity in real training environments. The next stage of framework development therefore requires prospective empirical validation in postgraduate medical education.

7.1 Recommended study design

The first empirical validation should ideally use a multicentre longitudinal design rather than a single cross-sectional survey.

A feasible study could recruit residents, clinical postgraduate students, or other postgraduate medical trainees early in a defined training period across multiple programmes that vary in specialty, institutional setting, and research infrastructure.

Baseline measurement should occur before, or as close as possible to, the principal resource-exposure period. Relevant pre-resource characteristics may include previous research experience, prior publications, formal methods training, specialty, stage of training, academic qualifications, and prespecified career intentions.

Research-resource conditions should then be measured prospectively or repeatedly during follow-up. These measurements should include, where feasible, mentorship access, protected research time, funding, data and infrastructure access, methodological and statistical support, and collaboration opportunities.

Prespecified scholarly outcomes should be measured after an appropriate follow-up period.

This temporal structure would provide substantially stronger support for causal interpretation than retrospective measurement of resources and outcomes at a single time point.

7.2 Multicentre measurement strategy

A multicentre study is particularly important because research opportunity is likely to vary both within and between training programmes.

Resource measurement should therefore distinguish trainee-level access from programme-level provision. For example, a programme may offer a statistical consultation service while only a subset of trainees actually obtain timely access to it.

Where possible, participant-reported measures should be complemented by administrative indicators such as scheduled research time, mentor assignment, research-funding availability, database permissions, methodological consultation use, and participation in structured research programmes.

Measurement invariance should also be examined when instruments are applied across specialties or institutions. A nominally identical resource measure may not have the same practical meaning in laboratory-intensive, surgical, epidemiological, or predominantly clinical research environments.

7.3 Benchmark validation

The adequacy benchmark R* should be specified before estimation and justified independently of the observed scholarly outcomes.

An initial multicentre study could examine several plausible benchmark definitions, including thresholds derived from accreditation expectations, expert consensus, programme policy, or prespecified percentiles within comparable specialty and training-stage strata.

The primary analysis should use one prespecified benchmark, accompanied by sensitivity analyses across alternative plausible thresholds.

Strong dependence of substantive conclusions on a narrow benchmark choice would indicate that the resulting RARL-R estimates are not yet sufficiently stable for educational decision-making.

7.4 Validation targets

Empirical validation should address several questions beyond whether RARL-R can be numerically calculated.

First, construct validity should examine whether the measured resource domains correspond to theoretically and educationally meaningful aspects of the research environment.

Second, predictive and calibration performance should evaluate whether estimated conditional expected outcomes are consistent with subsequent observed scholarly outcomes at appropriate aggregate levels.

Third, causal credibility should be strengthened wherever possible using longitudinal resource changes, programme reforms, natural experiments, or quasi-experimental contrasts.

Fourth, robustness should be evaluated across alternative outcome definitions, benchmark choices, model specifications, and assumptions regarding clustering and interference.

Fifth, educational interpretability should be assessed by determining whether estimated opportunity deficits correspond to resource domains that programme leaders and trainees recognize as actionable constraints.

7.5 Intended outputs of an empirical RARL-R study

A first empirical study should not aim to produce a league table of institutions or a ranking of individual trainees.

More informative outputs would include the distribution of resource-attributable expected output gaps, the proportion of trainees experiencing deficits in particular resource domains, variation across programmes, and changes in estimated gaps following resource-remediation initiatives.

For example, a programme could show relatively small deficits in funding but substantial deficits in protected research time and methodological support. Such a profile would imply a different educational response from a programme in which data access or mentorship represents the dominant constraint.

The empirical value of RARL-R therefore lies in whether it helps identify potentially modifiable research-training conditions and evaluate whether changes in those conditions are followed by corresponding changes in expected scholarly development.

Real-world longitudinal validation, rather than additional simulation alone, is the necessary next step in determining whether RARL-R can become a useful framework for postgraduate medical education research and programme evaluation.

8. Implications for Postgraduate Medical Education Research, Practice, and Policy

RARL-R is intended to complement, rather than replace, conventional evaluation of scholarly achievement and research quality. Its central contribution is to separate two questions that are often implicitly combined in postgraduate medical education: what scholarly work did a trainee produce, and under what research-opportunity conditions was that work produced?

The first question concerns observed achievement. The second concerns the extent to which expected scholarly development may differ under a specified and feasible improvement in the research environment. These questions are related but conceptually distinct. Two trainees may be assessed against the same scholarly expectations while having substantially different access to mentorship, protected research time, methodological support, data, infrastructure, or collaboration. RARL-R provides a formal framework for examining the second question without replacing direct assessment of the first.

8.1 Implications for medical education research

For medical education research, RARL-R offers a way to move from documenting barriers to specifying explicit and testable resource-remediation contrasts.

Existing research on resident and postgraduate scholarly activity has identified multiple barriers and facilitators, including mentorship, clinical workload, protected research time, research infrastructure, methodological expertise, and funding. Programme evaluations have also examined whether interventions targeting these factors improve scholarly participation or productivity [1–3,8].

RARL-R complements these approaches by defining a different target quantity: the expected difference in scholarly outcome associated with moving specified research-resource domains from their observed levels to a prespecified feasible adequacy benchmark, conditional on measured baseline characteristics.

This distinction may be particularly useful when educational researchers wish to move beyond statements that a resource is "associated with" research productivity. By requiring explicit definition of the resource intervention, adequacy benchmark, target population, outcome, and relevant baseline covariates, RARL-R encourages research questions that are closer to actionable programme decisions.

For example, rather than asking only whether protected research time is associated with publication, an RARL-R study could ask how expected scholarly output would differ if trainees currently receiving less than a defensible amount of protected time were brought to that level while other measured baseline characteristics were held fixed.

The framework may therefore provide a bridge between observational research on educational environments and intervention-oriented programme evaluation. It does not remove the need for strong study design; rather, it makes the intervention contrast and target estimand explicit before statistical analysis.

8.2 Implications for postgraduate research training programmes

For programme leaders, the principal value of RARL-R is diagnostic rather than evaluative.

Traditional scholarly metrics can identify trainees or programmes with relatively high or low research output, but they provide limited information about which modifiable features of the training environment may contribute to those differences. RARL-R redirects attention toward the resource conditions under which scholarly activity occurs.

A programme could, for example, show relatively adequate access to funding and clinical data but substantial deficits in protected research time and methodological support. Another programme may provide adequate research time but have limited supervisory capacity. Although these programmes could have similar average publication rates, the educational responses required would be different.

RARL-R could therefore support a structured process in which programmes:

1. identify research-resource domains that fall below defensible adequacy thresholds;
2. estimate the expected scholarly gap associated with those deficits;
3. prioritise potentially remediable constraints;
4. implement targeted changes; and
5. evaluate whether resource improvement is followed by changes in scholarly development.

This interpretation is intentionally different from assigning an "adjusted productivity score" to individual trainees. A high estimated RARL-R value does not imply that a trainee was entitled to produce a particular number of publications, nor does it establish what that individual would certainly have achieved under different circumstances. It indicates that, under the specified model and

assumptions, the observed research-resource profile differs meaningfully from a feasible remediation condition in terms of conditional expected scholarly outcome.

8.3 Implications for programme evaluation and resource allocation

At programme level, RARL-R should be interpreted as a distribution of potential remediation needs rather than as a single institutional performance score.

A programme with a moderately elevated average RARL-R may have widespread but relatively small deficits across many trainees. Another programme with the same average may provide adequate opportunities for most trainees while a smaller subgroup experiences severe deficiencies. These patterns have different educational implications despite producing the same programme mean.

Accordingly, programme-level reporting should, where sample size permits, include measures such as the median and interquartile range of estimated RARL-R, the proportion of trainees below specific resource adequacy thresholds, and domain-specific patterns of deficit.

Such reporting may help distinguish whether a programme's principal constraint relates to mentorship, clinical scheduling, methods support, data infrastructure, funding, collaboration, or combinations of these factors.

These findings could inform resource allocation, but RARL-R should not be used mechanically to determine funding decisions. Estimates remain dependent on measurement quality, benchmark specification, causal assumptions, model choice, and empirical overlap. Resource allocation decisions should therefore combine RARL-R estimates with qualitative evidence, programme priorities, feasibility, costs, and the educational importance of resource domains that may not be fully captured by scholarly output alone.

8.4 Implications for equity in postgraduate medical education

RARL-R also clarifies a particular dimension of educational equity.

Equal scholarly expectations do not necessarily imply equal opportunities to meet those expectations. Residency and postgraduate medical programmes may apply common requirements for research participation, presentations, thesis completion, or publication while providing heterogeneous access to the resources required to fulfil them.

RARL-R does not define fairness as identical resource allocation. Instead, it focuses on whether specified research opportunities fall below a defensible adequacy level and whether remediation of those deficits is expected to influence scholarly development.

This distinction between resource equality and resource adequacy is important. A trainee whose research environment already exceeds the adequacy benchmark is not assigned additional resource-attributable loss simply because another trainee has even greater resources. The framework therefore identifies potentially remediable insufficiency rather than treating every observed resource difference as an inequity.

At the same time, RARL-R addresses only one dimension of educational equity. Structural inequalities related to institutional selection, geography, socioeconomic circumstances,

discrimination, professional hierarchy, or broader workforce policy cannot be reduced to a resource-remediation estimand. These factors may shape both resource access and scholarly trajectories and require complementary qualitative, sociological, and policy-oriented approaches [9,10,21].

RARL-R should therefore be viewed as one quantitative tool within a broader programme of equity research rather than as a comprehensive measure of fairness in medical education.

8.5 Appropriate and inappropriate uses

The intended use of RARL-R is programme diagnosis, research-environment evaluation, and assessment of potentially modifiable educational conditions.

Appropriate applications may include comparing resource-deficit patterns across programmes, evaluating changes following introduction of protected research time or structured mentorship, identifying trainee groups with limited access to methodological support, and examining whether institutional interventions reduce estimated opportunity gaps over time.

By contrast, RARL-R should not be used as a stand-alone metric for trainee selection, promotion, certification, fellowship appointment, hiring, or disciplinary decisions. Nor should institutions be ranked solely according to mean RARL-R values. Such safeguards are consistent with broader principles of responsible and balanced research assessment [16–20].

Such applications would exceed what the estimand is designed to support and would risk converting an opportunity-sensitive educational framework into another individual performance metric.

The distinction is particularly important because the framework explicitly seeks to avoid attributing environmental constraints to personal deficiency. Its most defensible use is therefore to ask what aspects of the training environment may be improved, rather than to produce a revised hierarchy of trainees.

8.6 Relationship to the broader RARL research programme

RARL-R represents the resource-focused component of a broader Research Allocation and Resource Loss (RARL) research programme.

The present framework deliberately isolates a limited intervention: measured baseline trainee characteristics and the observation horizon are held fixed, while specified research-resource deficits are changed according to a feasible no-downgrade remediation policy. This restriction produces a clearer estimand and a more tractable empirical validation problem.

Other determinants of scholarly development—including task allocation, institutional incentives, geographical accessibility, career structure, and dynamic accumulation of research capital—may require separate models and should not be assumed to be captured by RARL-R.

For this reason, the present study should be interpreted as developing and testing one defined component of a broader research agenda rather than proposing a complete model of postgraduate academic development. At a broad conceptual level, the distinction between observed output and a feasible potential-output benchmark has an analogy with economic discussions of potential gross

national product [22]. This is an analogy only: RARL-R is defined here as an individual-conditional educational estimand and is not a direct adaptation of a macroeconomic output-gap measure.

9. Limitations and Future Research

RARL-R is intended as a formal framework for defining and estimating research-opportunity gaps, but several limitations constrain its interpretation and current stage of application.

9.1 Limitations

First, RARL-R is inherently counterfactual and therefore model-based. The scholarly outcome that a trainee would have experienced under an alternative resource profile is never simultaneously observable with the outcome under the observed profile. Potential-outcomes notation does not, by itself, make the resulting estimate causal. A resource-attributable interpretation depends on credible study design, appropriate temporal ordering, adequate control of pre-resource confounding, sufficient empirical support for the remediation contrast, and valid measurement of both resources and outcomes.

In observational medical education data, these requirements may be difficult to satisfy. Factors such as prior research motivation, supervisor selection, institutional culture, career orientation, or informal access to research networks may influence both resource exposure and subsequent scholarly development. If such factors are incompletely measured, estimated RARL-R values may reflect residual confounding rather than resource-attributable differences. Similarly, counterfactual estimates may become unstable when the remediated resource profile lies outside the combinations commonly observed for comparable trainees. RARL-R should therefore be described as resource-attributable only when the identification assumptions are sufficiently defensible; otherwise, a more cautious resource-associated interpretation is warranted.

Second, the definition and measurement of research opportunity remain substantive challenges. Mentorship, protected research time, methodological support, data access, funding, and collaboration are multidimensional constructs whose formal availability may differ from practical access or effective use. Measurement error in these domains may propagate directly into the remediation profile and estimated counterfactual outcomes. Moreover, a resource that appears equivalent across institutions may function differently depending on specialty, programme structure, clinical workload, or local research infrastructure.

The adequacy benchmark R* introduces an additional source of uncertainty. There is no universally correct threshold separating adequate from inadequate research opportunity. Benchmark selection inevitably contains educational and normative judgement, even when informed by programme standards, accreditation requirements, expert consensus, or empirical distributions. RARL-R estimates should therefore not be treated as benchmark-free quantities. The primary benchmark should be prespecified and justified independently of the observed outcome wherever possible, with sensitivity analyses conducted across alternative plausible thresholds.

Third, scholarly development is multidimensional and difficult to represent with a single outcome. Publication counts capture only one component of research activity, while composite scholarly-contribution scores necessarily embed choices about which outputs should be valued and

how they should be weighted. Journal-level metrics such as Journal Impact Factor are particularly unsuitable as individual-level indicators of scholarly quality.

For this reason, empirical validation should compare results across several transparent, prespecified outcomes rather than rely on a single opaque productivity index. The equivalent research-time transformation requires an even stronger measurement assumption: the underlying outcome must have a meaningful zero and an interpretable ratio scale. When these requirements are not satisfied, the resource-attributable expected output gap should remain the primary estimand and should not be expressed in equivalent research years.

Fourth, resource-outcome relationships may be nonlinear, heterogeneous, and interactive. More of a particular resource does not necessarily produce a proportional improvement in scholarly output. Additional protected time may have little effect in the absence of adequate mentorship or data access, whereas methodological support may be more valuable for some research projects or trainee profiles than others. Some resources may also reach saturation points or generate diminishing returns.

The positive-part operator ensures that an estimated decrease following remediation is not labelled as resource-attributable loss, but this mathematical convention does not establish that all resource effects are beneficial or monotonic. Empirical studies should therefore permit scientifically plausible nonlinearities and interactions, examine heterogeneity where sample size permits, and explicitly report situations in which a proposed remediation is estimated to have little, no, or potentially adverse effect.

Fifth, interference may be particularly important in postgraduate medical research environments. Supervisors, statistical services, databases, laboratory facilities, and collaboration networks are commonly shared. Increasing access for one trainee may improve opportunities for others through knowledge spillovers, but it may also reduce availability when resources are capacity constrained. Such processes violate simple individual-level no-interference assumptions.

Future empirical applications may therefore require programme-level clustering, multilevel modelling, cluster-level interventions, or network-aware approaches. The appropriate strategy will depend on whether the intervention acts primarily at the individual, supervisory-team, programme, or institutional level.

Sixth, RARL-R is vulnerable to misuse if interpreted as an adjusted measure of personal merit. The framework does not estimate innate research ability, deservingness, motivation, or the amount of scholarship that an individual “should” have produced. It should not be used as a stand-alone metric for admissions, selection, promotion, hiring, certification, or punitive performance management.

Such uses would invert the purpose of the framework by converting an environmental diagnostic into another individual ranking instrument. They could also create incentives to manipulate resource measures, benchmarks, or outcome definitions. The more defensible application is programme-level diagnosis and evaluation of potentially remediable research conditions, accompanied by transparent reporting of uncertainty.

Finally, RARL-R has not yet been empirically validated in real postgraduate medical education populations. The present Monte Carlo study demonstrates internal recoverability under a

controlled and correctly specified data-generating mechanism. It does not establish construct validity, causal validity, calibration, transportability, or educational usefulness in actual training programmes. This distinction is central to the interpretation of the present study.

9.2 Priorities for future research

The immediate research priority is prospective, multicentre longitudinal validation in postgraduate medical education. Such studies should recruit trainees before or early in the principal resource-exposure period, measure relevant baseline characteristics before those resources are experienced, repeatedly assess research-resource conditions during training, and ascertain prespecified scholarly outcomes after sufficient follow-up.

Future validation should address at least five methodological questions.

First, studies should evaluate the reliability and construct validity of the proposed research-resource domains and determine whether measurements are comparable across specialties and institutions.

Second, the adequacy benchmark should be examined empirically and conceptually. Analyses should assess whether substantive conclusions remain stable across alternative plausible thresholds and whether benchmarks require specialty- or training-stage-specific adaptation.

Third, estimator robustness should be evaluated under realistic departures from ideal conditions. Future simulation and empirical analyses should investigate model misspecification, nonlinear resource effects, measurement error, limited positivity, missing data, clustering, and interference.

Fourth, causal credibility should be strengthened wherever possible through programme reforms, natural experiments, stepped implementation of research-support interventions, or other quasi-experimental designs. Such designs would provide more informative tests of whether changes in the identified resource domains are followed by changes in scholarly outcomes than cross-sectional associations alone.

Fifth, validation should assess whether RARL-R produces actionable information beyond conventional programme-level measures. A useful framework should not merely generate statistically distinct numbers; it should help identify remediable constraints, discriminate between different patterns of research-resource deficit, and support evaluation of whether targeted interventions reduce those deficits over time.

Future work may also investigate decomposition of the total resource-attributable gap across interacting resource domains. For example, Shapley-value-based decomposition could be explored as a descriptive allocation of model-based contributions, although such decompositions should be interpreted cautiously because statistical attribution is not necessarily equivalent to a mechanistic causal contribution [23].

Most importantly, empirical validation should permit the framework to fail. If well-designed longitudinal or quasi-experimental studies find that remediation of proposed resource deficits produces little meaningful difference in scholarly development, the corresponding RARL-R estimates

should be small and the theoretical claims should be revised accordingly. Failure to reproduce meaningful resource-attributable gaps under stronger designs should count as evidence against the framework rather than as justification for post hoc redefinition of research "potential".

10. Conclusion

Scholarly development in postgraduate medical education is shaped not only by trainee characteristics but also by the research environments in which training occurs. Differences in mentorship, protected research time, methodological support, funding, data access, infrastructure, and collaboration may therefore contribute to differences in observed scholarly outcomes.

RARL-R provides a counterfactual framework for quantifying this specific dimension of research opportunity. Rather than treating observed output as a direct measure of individual capability, the framework compares conditional expected scholarly outcomes under the observed research-resource profile with those under a prespecified, feasible no-downgrade remediation of resource deficits. The resulting resource-attributable expected output gap is the primary estimand; where measurement assumptions permit, it may additionally be expressed as equivalent research time to aid interpretation.

RARL-R is not intended to estimate innate research potential, prescribe how much a trainee should have produced, or provide an adjusted ranking of individuals or institutions. Its intended role is diagnostic and evaluative: to make potentially remediable features of the research-training environment explicit and to support investigation of whether improving those conditions is associated with meaningful changes in scholarly development.

The present study establishes the conceptual and mathematical formulation of RARL-R, specifies its identification and measurement requirements, and demonstrates estimator recoverability under a controlled Monte Carlo setting. These results do not constitute empirical validation in real postgraduate medical education populations. The value of RARL-R will ultimately depend on prospective multicentre longitudinal and quasi-experimental studies that test its measurement validity, causal assumptions, benchmark sensitivity, robustness, and educational usefulness. The framework should be retained, modified, or rejected according to how well its prespecified resource-remediation contrasts are supported by real-world evidence.

Data and Code Availability

No human-participant, patient, or external observational data were used in this study. All numerical results were generated from synthetic data under the prespecified data-generating mechanism described in Supplementary Material S1.

The complete code used to generate the synthetic datasets, implement the RARL-R estimation procedure, reproduce the Monte Carlo analyses, and generate Table 4 and Figures 2–3 is provided in Supplementary Code S1.

The synthetic-data design and analysis code are provided to support reproducibility of the methodological results reported in this article. No restrictions related to participant confidentiality or clinical data access apply because no real individual-level data were analysed.

References

1. Laupland KB, Edwards F, Dhanani J. Determinants of research productivity during postgraduate medical education: a structured review. BMC Med Educ. 2021;21:567. doi:10.1186/s12909-021-03010-1.
2. Bakhshi SK, et al. Barriers to research productivity amongst postgraduate trainees: results from a survey of 333 medical and surgical residents. Postgrad Med J. 2023;99:1182–1188. doi:10.1093/postmj/qgad062.
3. Wood W, et al. Graduate medical education scholarly activities initiatives: a systematic review and meta-analysis. BMC Med Educ. 2018;18:318. doi:10.1186/s12909-018-1407-8.
4. Merton RK. The Matthew effect in science. Science. 1968;159(3810):56–63. doi:10.1126/science.159.3810.56.
5. DiPrete TA, Eirich GM. Cumulative advantage as a mechanism for inequality: a review of theoretical and empirical developments. Annu Rev Sociol. 2006;32:271–297. doi:10.1146/annurev.soc.32.061604.123127.
6. Mudrak J, Zabrodska K, Machovcova K. Crafting cumulative advantage: a systemic approach to the career development of highly productive researchers. Minerva. 2026;64:191–216. doi:10.1007/s11024-024-09567-7.
7. Bol T, de Vaan M, van de Rijt A. The Matthew effect in science funding. Proc Natl Acad Sci U S A. 2018;115:4887–4890. doi:10.1073/pnas.1719557115.
8. Rosenbloom JL, et al. The effects of research & development funding on scientific productivity: academic chemistry, 1990–2009. PLoS One. 2015;10:e0138176. doi:10.1371/journal.pone.0138176.
9. Deem R, Case JM, Nokkala T. Researching inequality in higher education: tracing changing conceptions and approaches over fifty years. High Educ. 2022;84:1245–1265. doi:10.1007/s10734-022-00922-9.
10. Mateos-González JL, Wakeling P. Exploring socioeconomic inequalities and access to elite postgraduate education among English graduates. High Educ. 2022;83:673–694. doi:10.1007/s10734-021-00693-9.
11. McAlpine L, Amundsen C, editors. Doctoral education: research-based strategies for doctoral students, supervisors and administrators. Dordrecht: Springer; 2011. doi:10.1007/978-94-007-0507-4.
12. McAlpine L, Amundsen C. Identity-trajectories of early career researchers: unpacking the post-PhD experience. London: Palgrave Macmillan; 2018. doi:10.1057/978-1-349-95287-8.
13. Rubin DB. Estimating causal effects of treatments in experimental and observational studies. ETS Res Bull Ser. 1972;1972:i–31. doi:10.1002/j.2333-8504.1972.tb00631.x.
14. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26:20–36. doi:10.1002/sim.2739.
15. Hernán MA, Robins JM. Causal inference: what if. Boca Raton (FL): Chapman & Hall/CRC; 2020. Updated online edition 2024.

16. Hicks D, et al. Bibliometrics: the Leiden Manifesto for research metrics. Nature. 2015;520:429–431. doi:10.1038/520429a.
17. Declaration on Research Assessment. San Francisco Declaration on Research Assessment. 2013 May 16. Available from: https://sfdora.org/read/.
18. Coalition for Advancing Research Assessment. Agreement on Reforming Research Assessment. 2022 Jul 20.
19. Declaration on Research Assessment. Guidance on the Responsible Use of Quantitative Indicators in Research Assessment. 2024. doi:10.5281/zenodo.13844873.
20. Denissen JJA, Sijtsma K, van der Aalst WMP. Towards a more informed and balanced use of scientific performance metrics. Res Eval. 2026;35:rvag023. doi:10.1093/reseval/rvag023.
21. Wapman KH, Zhang S, Clauset A, Larremore DB. Quantifying hierarchy and dynamics in US faculty hiring and retention. Nature. 2022;610:120–127. doi:10.1038/s41586-022-05222-x.
22. Okun AM. Potential GNP: its measurement and significance. In: Proceedings of the Business and Economic Statistics Section. Washington (DC): American Statistical Association; 1962. p.98–104.
23. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the theory of games II. Princeton (NJ): Princeton University Press; 1953. p.307–317.

Table 1. Mathematical properties of RARL-R for opportunity-sensitive interpretation.

Property	**Interpretation**
Non-negativity	RARL-$R_i \geq 0$ by construction.
Zero-loss consistency	If $\mu_i(R_i\dagger)=\mu_i(R_i)$, then RARL-$R_i=0$.

Boundedness	If $\mu_i(R_i)>0$ and $\mu_i(R_i\dagger)\geq\mu_i(R_i)$, then $0\leq RARL\text{-}R_i\leq T$.
Positive scale invariance	Replacing Y by cY for $c>0$ leaves the ratio $\mu_i(R_i)/\mu_i(R_i\dagger)$, and therefore RARL-R, unchanged.
No-downgrade benchmark consistency	If all observed opportunity components already meet R*, then $R_i\dagger=R_i$ and $RARL\text{-}R_i=0$.
Monotonicity under productive remediation	If additional opportunities weakly increase expected output, reducing an opportunity deficit cannot increase RARL-R.

国家高端交流平台

Table 2. Candidate dimensions of institutional research opportunity structures for empirical RARL-R studies.

Domain	Recommended indicators
Mentorship	Scheduled one-to-one supervision time; mentor expertise contextualized by field; number of concurrently supervised researchers; availability of project-level feedback.
Protected time	Hours or proportion of programme time formally protected for research; frequency with which protected time is interrupted by teaching, clinical, administrative or service duties.
Funding	Researcher-accessible project funds; small-grant availability; support for core facilities, data collection, software, publication and conference participation.
Data & infrastructure	Access to discipline-relevant datasets, archives, laboratories, computational platforms, software, core facilities and equipment; waiting time and usability.
Methodological support	Statistics, qualitative methods, computational methods, data management, laboratory technical support and research-governance assistance.
Collaboration/network	Cross-institutional collaboration access, seminar/meeting exposure, expert consultation and referral pathways to complementary expertise.

国家高端交流平台

Table 3. Identification conditions for interpreting RARL-R as opportunity-attributable rather than opportunity-associated loss.

Condition	**Why it matters**
Consistency	The observed outcome under the observed opportunity vector equals the corresponding potential outcome.
Conditional exchangeability	After conditioning on measured pre-opportunity X, there are no unmeasured common causes of opportunity exposure and potential scholarly output.
Positivity	For covariate patterns used in the analysis, there is adequate empirical support for the relevant opportunity levels or remediation contrasts.
Temporal ordering	Key baseline covariates precede the opportunity exposure whose effect is being estimated; variables caused by opportunities are not naively adjusted as confounders.
Measurement validity	Opportunity and outcome scales are reliable, comparable across centres and sufficiently invariant across disciplines/stages.
Interference control	Shared mentors, collaborative networks and institutional spillovers may violate no-interference assumptions; clustering or network-aware designs may be needed.

国家高端交流平台

Table 4. Monte Carlo recovery of true RARL-R across 200 replications per sample size.

n	Mean true RARL-R (years)	Counterfactual RMSE (years)	Counterfactual r	Crude RMSE (years)	Crude r
200	0.886	0.209	0.938	0.603	0.628
500	0.882	0.132	0.976	0.595	0.638
1000	0.884	0.092	0.988	0.596	0.642

国家高端交流平台

Table 5. Recommended design for the first multicentre empirical validation of RARL-R in postgraduate medical education.

Element	Recommended first-study design
Population	Residents, clinical postgraduate students, and other postgraduate medical trainees recruited from multiple institutions and specialties; stratify by training pathway and specialty rather than pooling unlike groups uncritically.
Baseline X	Prior publications/projects, quantitative-method training, degree pathway, discipline, stage of study, prior research time, stated research intention, and other pre-resource variables justified by a causal diagram.
Resource R	Mentorship, protected time, funding, data/infrastructure, methods support and collaboration measured repeatedly where feasible.
Primary outcome	Prespecified non-negative scholarly-contribution measure over a fixed horizon; also report domain-specific outcomes separately.
Benchmark R*	Predefined adequate-opportunity thresholds or stage-/discipline-specific thresholds; perform sensitivity analyses across plausible benchmark values.
Primary estimator	G-computation with transparent model specification; cluster bootstrap by institution; centre-level random/fixed effects considered as design requires.
Validation	Convergent validity with independently measured opportunity access; known-groups comparisons; calibration of μ; sensitivity to benchmark, outcome definition and unmeasured confounding.
Ethical use	Report distributions and programme-level deficits; do not publish individual “potential” labels or use RARL-R for punitive ranking.

国家高端交流平台