Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Participation bias and false alarms in resistant-carriage surveillance: a simulation study

Chaokun Hong

chaokun.hong@ndm.ox.ac.uk

University of Oxford

Research Article

Keywords: antimicrobial resistance, carriage surveillance, participation bias, simulation study, model calibration, identifiability, infection control

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11295092/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Participation bias and false alarms in resistant-carriage surveillance: a simulation study

Chaokun Hong[1,2*]

1 Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

2 Oxford University Clinical Research Unit, National Hospital for Tropical Diseases, Dong Anh, Hanoi, Vietnam

* Corresponding author: Chaokun Hong; chaokun.hong@ndm.ox.ac.uk

Article type: Research article

Abstract

Background: Changes in resistant carriage among surveillance participants can reflect population change or changing participation. We examined whether explicit participation modelling and threshold calibration control false alarms, and whether good model fit identifies the population change.

Methods: Two mechanistic carriage structures generated synthetic surveillance at an exposed and a control site. The target was the population resistant-carriage difference from an unshocked counterfactual, averaged over 180 days. Joint sensitive/resistant measurements included participation selection, clustered sampling and diagnostic error. A finite library supported early-survey rules with and without temporal projection. Calibration, selection and evaluation used 17, 29 and 44 scenarios; evaluation used 3,000 replicates per scenario. Follow-ups examined participation coverage, a calibrated direct comparator, and 132,000 synthetic library fits at sensitive-carrier participation weights 0.8, 1.0 and 1.2 without recalibration. A deterministic diagnostic tested whether early observations identified perturbation duration.

Results: Participation shifts alone produced apparent resistant-carriage changes of approximately −4.0 to +3.7 percentage points. Under selection on resistant carriage alone, widening the library while calibrating against all original nulls reduced the dynamic rule's maximum false-alarm rate from 6.63% to 3.03% and increased mean power from 11.47% to 19.24%. At the widest library, mean power was 25.42% without projection and 11.02% for the calibrated direct comparator; corresponding maximum false-alarm rates were 2.97% and 2.77%. In the new stress test, sensitive-carrier weight 1.2 raised maximum false-alarm rates to 77.2% and 76.9% for the two model rules. Across all 44 scenarios at that weight, 99.93% of fits had small observed-summary residual distance, yet 63.13% combined small distance with population-target error exceeding 1 percentage point.

Conclusions: Modelling participation improved detection within the assumed selection process, but an unrepresented participation change invalidated the calibrated model rules despite close fit to observations. Low power also limits reassurance from a non-alert. Resistant-carriage monitoring requires explicit assessment of participation assumptions and external validation before these rules can guide infection-control decisions.

Keywords: antimicrobial resistance; carriage surveillance; participation bias; simulation study; model calibration; identifiability; infection control

Background

Antimicrobial resistance surveillance informs infection prevention and treatment policy, but changes in its measurements depend on how specimens are obtained. Population resistant-carriage prevalence, the proportion resistant among clinical isolates, and the incidence of resistant infection have different denominators. WHO guidance addresses representativeness and diagnostic access, and assessments of surveillance systems identify sampling and multiple-counting bias [1, 2]. Hospital data and simulations show how blood-culture utilisation affects estimated resistance proportions [3]. In an Indonesian referral hospital, an increase in hospital-origin resistant bloodstream infection incidence was interpreted in the context of increased blood-culture utilisation [4]. These findings motivate careful interpretation of observation processes; they do not supply empirical parameters for the carriage experiment reported here.

Carriage surveys can assess a population beyond those seeking clinical care, while still being affected by who agrees or is available to participate. If carriers and non-carriers participate at different rates, a change in relative inclusion can resemble a change in carriage. This is an outcome-dependent selection problem: observed proportions alone generally cannot separate the two processes without additional information or restrictions [5, 6, 7]. After a disruption to services or population access, a surveillance alert could therefore reflect biological change, altered participation, or both. Sampling strategies designed to detect a new resistance variant address a different objective and should be evaluated against that objective [8].

Mechanistic models can connect observations in a sampled window to a longer population trajectory. That connection creates two distinct requirements. The inferential model must represent plausible observation processes, and the threshold used to declare an increase must be calibrated under relevant null changes. A narrow model can give a good observed-data fit while assigning the wrong population explanation.

Increasing the threshold may reduce false alarms at the cost of detecting fewer real changes. Simulation studies allow these trade-offs to be examined with the population truth known [9, 10].

We studied decisions about a 180-day increase in population resistant carriage using two simplified mechanistic structures. We separated participation coverage in a finite prediction library from coverage of the null scenarios used to calibrate decisions, compared temporal projection with a matched rule without projection and a calibrated direct comparator, and examined whether fit diagnostics exposed an unrepresented participation mechanism. The study evaluates conditions under which a surveillance interpretation fails; it does not estimate pathogen-specific transmission parameters or validate a clinical alert system.

Methods

Aim, estimand and simulation design

We organised the study around aims, data-generating mechanisms, estimands, methods and performance measures, following the ADEMP reporting framework [10]. All data were synthetic. One site experienced a specified perturbation and a second remained unshocked. The population target was the signed difference in resistant carriage between the exposed site's trajectory and its unshocked counterfactual, averaged over integer days 0–179 and expressed in percentage points (pp). Stable counterfactual behaviour and no spillover were built into the experiment. A DiD interpretation requires a valid counterfactual trend; that assumption is supplied here by the simulated unshocked site [11].

Repeated cross-sections were collected during baseline days −90 to −1, early days 0–89 and late days 90–179. The direct estimator was the equal-weight average of the early and late difference-in-differences (DiD). Consequently, the baseline site contrast entered with total weight one. A one-sided decision declared an increase or remained undetermined. The target horizon was fixed at 180 days, including scenarios with longer carriage duration.

Synthetic population trajectories

Both structures started from sensitive-carriage prevalence S=0.56 and resistant-carriage prevalence R=0.14. Structure A allowed simultaneous sensitive and resistant carriage; its daily joint-state probabilities were the products of the two marginal carriage probabilities. Structure B allowed mutually exclusive uncolonised, sensitive and resistant states, with replacement between carriage states and no true co-carriage. Baseline

acquisition rates were solved so that each structure matched the same starting prevalences. These structures provided two synthetic sensitivity settings with the same baseline prevalence (Fig. 1a–b). S and R denote carriage type, not susceptible and recovered immune states; figures use Cs and Cr to make that distinction explicit. Carriage can be reacquired after clearance.

Perturbations increased acquisition pressure for both types or for the resistant type only, for a specified duration. Parameters governing carriage duration, external acquisition, replacement and resistant-carriage clearance varied over declared ranges (Additional files 1–2). Separate changes in care attendance and testing operated in the clinical observation process. Complete equations, parameter definitions and ranges are given in Additional file 1 and the code available on request. These are ordinary differential equation models of acquisition and clearance, with replacement in structure B. Matching baseline equilibrium, solving amplitudes for specified population targets, and fitting the finite library to simulated observations are distinct operations. None is an empirical fit to a real carriage time series. The parameter ranges define the synthetic experiment.

We also examined whether different perturbation amplitudes could produce the same 180-day target. Twelve curves crossed two structures, three carriage durations and two perturbation types. Roots were sought over an amplitude interval of 0–20 using nested grids, sign-change refinement and checks near stationary points. Every detected root was retained and assigned its own branch identifier. Normal-approximation budget calculations for each branch were checked by actual cluster simulation for the two-branch example.

Participation, joint measurement and cluster sampling

Let w be the participation weight of a resistant carrier relative to a non-carrier weight of 1. At a fixed population prevalence R, the participant prevalence is $Rw/(1-R+Rw)$. Baseline weights were 1.0 or 0.8 in the prediction library, and post-baseline changes were applied only at the exposed site. Participation changes such as ±0.30 are absolute changes in this relative weight. The simulation code parameterises the same quantity as $w=1-phi$.

Participation weights were applied to each day's complete joint sensitive/resistant state table, followed by normalisation, before averaging over sampling days. This correctly allowed selection on R to change observed S in the mutually exclusive structure. Completed observations were uniformly distributed across

dates within each window, corresponding to equal daily completion quotas or equivalent calendar-time weights. The original experiment allowed selection on R status only. The additional sensitivity experiment below relaxed this restriction.

Each independent cluster contained 20 participants. Cluster counts followed a Dirichlet–multinomial sampling model, a distribution used for overdispersed multinomial counts [12]. In our implementation, a Dirichlet draw generated the latent category composition and a multinomial draw generated participant states. Sensitive and resistant tests were then applied to the same participants. Test errors were independent conditional on true state. Sensitivity was 0.90 for each test; specificity was 0.95 for sensitive and 0.97 for resistant carriage. These accuracies were known to the estimator. True co-carriage was absent in structure B, but false-positive co-detection remained possible.

Design effects of 1, 2 and 4 referred to the observed resistant indicator after test error. The latent intracluster correlation was chosen to achieve the requested observed design effect; the sensitive indicator could have a different design effect. Estimation used empirical cluster means and their between-cluster covariance, followed by linear correction for test sensitivity and specificity [13]. Corrected estimates retained their full numerical range. The estimator used observed cluster means and covariances. Direct contrasts used a Satterthwaite t reference computed from the estimated variance components. Approximate t inference with an effective number of degrees of freedom has a classical basis for combining independent variance estimates; the particular plug-in calculation used here is specified in the code available on request [14].

Decision rules and the prediction library

Four candidate designs used an early survey, a late survey, surveys with two scheduled opportunities to act, or clinical validation plus a late survey. Each had a dynamic version and a version without temporal projection. All were compared with the direct two-window estimator. A common budget ceiling of 32,000 synthetic units applied; survey unit costs were 1, 3 or 8, with only whole clusters recruited. Baseline collection used one third of the survey budget. The remainder was allocated to the chosen post-baseline window or divided between both windows.

The library contained 742 population-trajectory templates, each crossed with two baseline participation weights and five participation changes, producing 7,420 rows. It combined a fixed parameter grid,

additional construction draws and the retained calibration branches. The early rule received two observed DiD summaries and their estimated covariance, together with the fixed library (Fig. 1d). Daily generating trajectories, late observations and generating parameters remained hidden. Fitting minimised distance over the finite entries; it did not continuously estimate a parameter vector from an observed time series.

For each library row, a squared Mahalanobis distance compared its predicted observed-window sensitive/resistant change with the measured change, weighting the discrepancy by the estimated covariance [15]. The decision score was the square root of the non-negative difference between the minimum distance among non-positive-target rows and the minimum over all rows. Thus a high score indicated a poor fit of every available non-positive population-change template relative to the best fit. The best-fitting row minimised that distance. A finite-library compatibility range contained targets of all rows within the chi-squared 95th-percentile tolerance with two degrees of freedom of the minimum. This range is a descriptive summary of the library, not a validated confidence interval or posterior interval. Fitting selected a synthetic trajectory and its target; it did not establish unique biological-parameter identification.

The temporal-projection ablation retained the same observed-window predictions and covariance. It replaced the missing window's population change by the observed window's population change when defining the 180-day target and its null subset. The ablation thus retained model-derived observed-window predictions. Two-look rules calibrated the maximum of the early prediction score and final direct-estimator score at the two scheduled analyses.

The validation design reserved one quarter of the budget for samples from both cultured and non-cultured clinical-attendee strata. Its estimated attendee signal provided a prespecified positive corroboration gate for the late-survey rule. Available savings could move forward in time but could not alter baseline recruitment retrospectively. A clinical reference sampled attendees and compared its signal with a nominal threshold. A calibrated version used the same null-tail calibration principle. Clinical signals served as attendee-based decision proxies for the population target. Different designs shared prefixes of generated cluster samples where their collection plans overlapped, enabling paired comparisons without supplying uncollected observations to a rule. Paired comparisons retained the covariance induced by these common random numbers [10, 16].

Information available before the late window

An explanatory follow-up examined what the early observations could identify (Fig. 2). The illustrative case was the first evaluation scenario by ID with structure B, a common perturbation and target +1 to +5 pp, using S weight 1 and the first replicate of the additional fitting experiment. The choice preceded inspection of its curves. We reconstructed its early estimates, selected entry and every compatible trajectory, without treating the daily generating values as observed data.

For a deterministic duration diagnostic, we held this scenario's remaining parameters and perturbation amplitude fixed and varied duration from 90 to 180 days in five-day increments. We compared the full baseline/early joint-state observation probabilities and the 180-day target. Because the early window ends at day 89, every duration leaves the perturbation active throughout that window. We checked the endpoint trajectories with a separately implemented fixed-step Runge–Kutta integrator. This diagnostic followed inspection of the original study and is an explanatory counterexample, not an independently designed evaluation or a new forecasting rule.

Calibration, selection and evaluation

Separate parameter draws and recorded seeds defined 17 calibration scenarios, 29 selection scenarios and 44 evaluation scenarios. Each phase included stable, constant-informative and changing-informative participation. Null scenarios included service or participation changes with zero population change, and dynamic paths whose mean target was zero despite changes within the two windows. The evaluation set comprised 14 null and 30 positive-target scenarios; 18 positive scenarios had targets between +1 and +5 pp and formed the primary power summary. The remaining 12 larger changes were retained in scenario-level results.

Calibration, selection and evaluation used 4,000, 1,000 and 3,000 replicates per scenario, respectively. Prediction thresholds were the maximum across calibration nulls of each scenario's 97.5th score percentile, with a small numerical increment. The calibration tail left a simulation margin while the scientific false-alarm criterion remained 5%. Nominal reference rules retained a one-sided 5% threshold. Thresholds were saved before selection and evaluation.

Within the dynamic and ablation candidate sets separately, selection maximised mean power over +1 to +5 pp scenarios among rules whose simultaneous 95% Monte Carlo upper bounds were at most 5% for every

selection null. Both selections chose the early-survey design. Choices were then fixed for the evaluation partition. The calibrated clinical reference was added with its algorithm specified before its evaluation estimates were inspected; it did not change the selected rules. The participation-width analysis followed examination of the original results and reused the evaluation set as a sensitivity experiment.

Participation-width and calibration-coverage analysis

We fixed the original early-survey rules, budgets, evaluation scenarios and observation seeds, and restricted the library to participation changes of 0, ±0.15 or ±0.30. The resulting libraries had 1,484, 4,452 and 7,420 rows, preserving all 742 population templates and both baseline weights. Two threshold-calibration policies were crossed with each width. The within-range policy used only original calibration nulls whose participation changes lay within the stated width. At widths 0 and ±0.15, this retained 12 nulls with no differential participation because the nonzero calibration changes were −0.25 and +0.20. At ±0.30, all 17 nulls were included. The all-null policy used the same 17 calibration nulls at every library width.

The evaluation set was unchanged under both policies. Consequently, narrow-library evaluations deliberately included participation changes outside the library's stated range. This crossing distinguishes an anticipated-drift calibration from a stress test in which drift was absent from both the narrow calibration set and the narrow library. Thresholds were saved before each follow-up evaluation, and the widest-library scores reproduced the original replicate-level scores. The selected observation design remained fixed.

Calibrated direct comparison and participation stress test

The original direct two-window reference used a nominal one-sided 5% threshold. We added a comparator using its unchanged direct score but the same calibration principle as the model rules: the largest scenario-specific 97.5th percentile across all 17 original nulls, plus a numerical increment. The threshold was fixed before recalculating its evaluation decisions. This follow-up used previously generated scores and previously inspected evaluation scenarios, so it is an exploratory comparison. It held the budget ceiling and calibration scenarios constant, although the direct design sampled both follow-up windows whereas the early designs concentrated follow-up recruitment in the first window. Paired differences used shared simulated observations.

We then generated 1,000 new replicates for each of the 44 evaluation scenarios under sensitive-carrier participation multipliers of 0.8, 1.0 and 1.2, yielding 132,000 synthetic fits. At the exposed site after

baseline, each joint-state probability received the product of its original R participation weight and the multiplier if S was present, followed by daily normalisation. Baseline and control-site S multipliers remained 1. The value 1.0 reproduced the original participation formula; 0.8 and 1.2 deliberately introduced a mechanism absent from the library. These values were chosen as sensitivity conditions, without empirical calibration. The full ±0.30 library, model thresholds and calibrated-direct threshold remained fixed. Biology, diagnostic accuracy and the surveillance budget were unchanged.

For every replicate, the early S/R contrast was fitted to the finite library. We recorded the fitted population target, its error against the known generating target, minimum observed-summary distance, and target compatibility range. A distance no larger than the chi-squared 95th percentile with two degrees of freedom was labelled a small residual for descriptive purposes; it is not a formal goodness-of-fit test after fitting. We also recorded the fraction combining a small residual with absolute population-target error above 1 pp. These diagnostics cover all 44 scenarios, including the 12 larger increases; false-alarm and power summaries retain the original 14 nulls and 18 small increases. New random seeds provide new simulation observations, while reuse of the same model families and scenario definitions makes this an internal sensitivity analysis rather than external validation.

Performance measures and uncertainty

False-alarm performance was the maximum declaration probability among the same 14 evaluation nulls. Power was the equal-weight mean across the same 18 evaluation increases of +1 to +5 pp; minimum and maximum scenario power were also retained. These summaries give equal weight to the specified scenarios. Clopper–Pearson binomial upper bounds were combined by a Bonferroni union bound across null scenarios to provide simultaneous 95% Monte Carlo bounds within each rule [17, 18].

Mean-power Monte Carlo intervals used the scenario-specific Bernoulli simulation variances and independent scenario replicates. Paired dynamic-minus-ablation differences used the within-replicate decision difference, retaining the pairing of observations. Bias, root-mean-square error, direct-estimator interval coverage, compatibility-range inclusion and structure classification were supporting measures. Monte Carlo intervals quantify repeated simulation error conditional on the fixed scenarios [9, 10].

Analyses and graphics were implemented in R 4.4.3; numerical trajectory integration used deSolve [19, 20]. Saved scripts, seeds, aggregate outputs and synthetic decision-score records support reproducibility.

Results

Participation alone can mimic a population change

At constant population resistant carriage of 0.14, changing the R-carrier participation weight by ±0.30 produced apparent changes of −3.77 to 3.47 pp from a baseline weight of 1.0 and −4.00 to 3.66 pp from a baseline weight of 0.8 (Fig. 1c; Additional file 1, Fig. S1). These shifts are of the same order as the small population increases targeted in the power analysis. The apparent sensitive-carriage prevalence stayed unchanged in structure A because the joint probabilities factorised, whereas it changed in structure B because sensitive and resistant carriage were mutually exclusive. Thus the information supplied by joint S/R measurements depended on the joint-state and selection assumptions.

Early fit does not resolve the unobserved trajectory

In the illustrative replicate, the selected library entry closely matched both early contrasts, with minimum squared distance 0.000135. The generating 180-day target was +1.00 pp and the selected target was +1.17 pp, but 2,738 compatible entries represented 428 distinct trajectories and targets spanning 0 to +21.84 pp (Fig. 2a,c). This single, rule-selected example illustrates the workflow; population-target recovery across all scenarios is reported below and in Additional file 1.

Duration remained unresolved even without sampling noise. All 19 durations continuing beyond the early observation window produced identical baseline/early joint observation probabilities. Their 180-day targets ranged from +0.814 pp at 90 days to +1.424 pp at 180 days (Fig. 2b). Thus, under these fixed remaining parameters, the early observation distribution did not identify the later target. The alternative solver agreed within 1.05×10^{-10} in prevalence. This restriction is separate from residual fit and cannot be removed by optimising duration against the same early information.

Calibration coverage changes the effect of library width

When calibration used only nulls inside the stated library width, the dynamic rule's maximum false-alarm rate declined from 30.37% at zero width to 3.03% at ±0.30; mean power fell from 24.28% to 19.24%. Narrow-library evaluation included participation drift absent from calibration, so its higher apparent power accompanied failed error control (Table 2; Additional file 1, Fig. S2).

Using all 17 original calibration nulls at every width changed the trade-off: widening the library reduced maximum false alarms from 6.63% to 3.03% while increasing mean power from 11.47% to 19.24%. No-

projection rates similarly changed from 6.50% to 2.97%, with power increasing from 17.97% to 25.42%. Even all-null calibration left the zero-width maximum false-alarm point estimate above 5%. Intermediate widths and uncertainty are retained in Table 2 and Additional files 1–2.

Held-out rule performance and temporal projection

At the original ±0.30 library width, the selected early dynamic rule had mean power 19.24% and maximum false-alarm rate 3.03%, with a simultaneous 95% Monte Carlo upper bound of 3.98%. The early no-projection rule had mean power 25.42% and maximum false-alarm rate 2.97%, with upper bound 3.90%. Its paired advantage over the dynamic rule was 6.176 pp (95% Monte Carlo interval 5.994 to 6.358). Thus temporal projection reduced detection in this fixed scenario set.

Applying all-null calibration to the direct two-window rule reduced its maximum false-alarm rate from 30.77% to 2.77% and its mean power from 33.79% to 11.02% (Fig. 3a–b; Additional file 1, Fig. S5). The calibrated direct rule's simultaneous false-alarm upper bound was 3.67%. Relative to this comparator, the early dynamic and no-projection rules gained 8.23 pp (95% Monte Carlo interval 7.95 to 8.50) and 14.40 pp (14.15 to 14.66) in mean power. Paired differences varied substantially across the 18 scenarios, including negative differences (Fig. 3c). The comparison supports a conditional benefit of joint model-based inference and early sampling under the original observation assumptions; comparison with the nominal direct reference alone would overstate this evidence.

The complete original comparison is retained in Additional file 1, Table S3. Late dynamic, two-look dynamic and validation-plus-late dynamic rules had mean powers of 17.52%, 13.99% and 12.49%. The calibrated clinical proxy had mean power 1.79% and maximum false-alarm rate 2.10%, compared with 28.51% and 80.57% for its nominal counterpart. These proxy decisions concern the population-carriage target and do not assess the usefulness of clinical surveillance for its own infection or treatment objectives.

All designs respected the 32,000-unit budget ceiling. Early surveys spent 31,680–31,960 units and direct two-window surveys spent 31,680–31,920. Clinical-reference expenditure was 5,182.8–8,618.4 units because available attendees limited recruitment. A common budget ceiling did not guarantee equal realised expenditure.

Detection differed by joint carriage structure and perturbation

The model rules did not dominate every subgroup. Early dynamic and no-projection mean power ranged from 1.78% and 2.12% in structure A with R-specific perturbations to 42.69% and 56.03% in A with common perturbations. In B with R-specific increases, calibrated direct reached 10.81%, above the two early rules' 9.82% and 9.99%. The four groups partition all 18 small-increase scenarios but also differ in duration, cost, clustering and participation; their means do not isolate a biological-structure effect. Complete group counts and means remain in Additional file 1, Table S5.

Unrepresented participation changes defeated calibration despite close fit

With no S-dependent participation change in the new simulations, maximum false-alarm point estimates were 4.1%, 4.0% and 1.6% for early dynamic, no-projection and calibrated direct rules, respectively. The simultaneous upper bounds were 6.09%, 5.97% and 2.99%. Thus the smaller 1,000-replicate rerun did not certify a 5% bound for the two model rules even under the identity condition. Mean powers were 19.29%, 25.79% and 11.02%, consistent with the original comparison.

At sensitive-carrier weight 1.2, maximum false-alarm rates rose to 77.2% for the dynamic rule and 76.9% without projection; their simultaneous upper bounds were 80.67% and 80.39% (Fig. 4; Table 3). The calibrated direct comparator remained at 1.9%, with upper bound 3.38%, but had mean power of only 8.27%. At weight 0.8, maximum false alarms were 3.8%, 11.1% and 9.5%, respectively. No rule was uniformly protected across these unrepresented participation conditions.

Close fit to observed summaries did not reveal the failure. At S weight 1.2, 99.93% of fits had small residual distance, while 63.13% of all fits combined small distance with population-target error greater than 1 pp. At the identity weight, the corresponding values were 99.80% and 66.59%; root-mean-square target error was 6.14 pp across all 44 scenarios. The median across scenario-specific median compatibility widths was 27.50 pp under the identity condition and 25.35 pp at weight 1.2. These broad ranges and errors show that the available summaries did not tightly identify population trajectories, even when observations were closely matched (Additional file 1, Table S7 and Fig. S4).

Supporting trajectory and observation checks

The bounded branch search retained 46 roots, including 34 nonzero roots, and demonstrated that the same 180-day target could arise from different temporal paths. Joint S/R data improved correct-structure

classification under the original selection assumption, but uniquely resolved structure in only 11.26% of positive-scenario replicates on average. Full root checks, branch-specific budgets and simulations, and classification results are retained in Additional file 1, sections S4 and S7.

Discussion

The interpretation of a resistant-carriage alert depended on the observation mechanism represented by the model. Under the original R-only participation process, including a wider participation range improved power and false-alarm performance together when thresholds were calibrated against the same nulls. Joint model-based early monitoring also detected more small increases than a direct two-window comparator calibrated on the same null scenarios. However, adding an unrepresented S-dependent participation change produced false-alarm rates approaching 80% for the two model rules. A claim of general robustness would therefore be unsupported.

The stress test explains why model fitting alone cannot establish population interpretation. The model rules use S as well as R. In the factorised structure, R-only selection leaves participant S unchanged, allowing S to help distinguish common biological changes from R-dependent participation. That restriction is informative, but it can be wrong. When S also affects participation, an observed shift can closely match a library trajectory while arising from a different population process. High fit quality and large target error can then coexist. The very wide compatibility ranges further show how little the early summaries resolve the underlying trajectories. Reporting fit residuals, fitted targets and the range of compatible targets together is more informative than reporting a best-fitting curve alone [6, 7].

Calibration and model coverage played complementary roles within the original mechanism. A narrow library omitted null explanations for some participation-induced shifts, so calibration had to compensate through a higher threshold. A broader library supplied those explanations and retained greater detection. The calibrated-direct comparison also shows why nominal references should not be used to judge power without examining their false alarms: its apparent power fell substantially after comparable null calibration. The remaining model advantage combined use of joint S/R information, library assumptions and earlier concentration of sampling effort. It cannot be attributed solely to fitting a mechanistic trajectory.

Temporal projection added no detection benefit on average in this experiment. The matched no-projection rule retained model-derived predictions and used the observed-window population change for the missing

window. Its greater power therefore does not demonstrate that models are unnecessary; it identifies a cost of this particular projection under limited observations and the specified target. The duration diagnostic identifies a specific information limit: an early survey ending before a perturbation stops cannot reveal its eventual stopping time. Continuous optimisation of the same early summaries cannot recover that information without additional restrictions. The diagnostic does not preclude estimation of other parameters under stronger assumptions. Later observations or defensible duration information could alter the projection trade-off. Detection also remained weak: the best of the three principal calibrated rules detected only about one quarter of the specified small-increase scenarios on average, and much less in some subgroups. An undetermined result must not be interpreted as evidence that population carriage was unchanged.

For infection-control surveillance, the practical implication is to document how observations were obtained and test the resulting assumptions. Invitations, completion by sampling stratum and time, refusals, clinical testing volume and changes in diagnostic access help describe the observation process [1, 3, 4]. Overall response counts alone cannot identify a participation ratio conditional on unobserved carriage. More representative validation samples or defensible external information would be needed to estimate that ratio; otherwise sensitivity or partial-identification analyses should express the unresolved selection [7, 21, 22]. The clinical proxy's poor performance for a population-carriage target does not diminish the value of clinical microbiology for guiding treatment or monitoring infection among patients. The target, denominator and sampling frame must remain explicit.

Several limitations constrain application. All trajectories and observations were synthetic, with no pathogen-specific empirical fitting. Two simplified carriage structures supplied both the construction library and evaluation framework; separate seeds and parameter draws do not provide external validation. Test accuracies were known, diagnostic errors conditionally independent, and sampling used equal daily completion with fixed cluster size. The added S-dependent weights examine one omitted observation mechanism, without establishing that 0.8–1.2 is a clinically plausible range or covering other selection pathways. The finite grid, bounded root search and small scenario family cannot establish uniform error control. The width, calibrated-direct, fitting and duration diagnostics were follow-ups using previously inspected scenario definitions. The duration counterexample holds other parameters fixed within one scenario; it does not classify identifiability for all possible carriage models or observation designs. Monte

Carlo intervals quantify repeated simulation error within those definitions, not uncertainty about transfer to a real pathogen, service or population [9, 10, 23].

An empirical evaluation would require repeated carriage observations with population denominators, sampling and participation information, joint measurements and setting-specific diagnostic performance. It should assess both forecast or target accuracy and decision errors on observations withheld from fitting. The present study provides a reason to demand that evidence: a model can explain the available surveillance measurements closely while supporting the wrong population conclusion.

Conclusions

Participation modelling and null calibration improved resistant-carriage monitoring within the assumed observation process. An unrepresented S-dependent participation change nevertheless caused severe false alarms despite close fit to observed summaries. Low detection also limits interpretation of non-alerts. These simulation findings support explicit participation sensitivity analysis and setting-specific external validation before model-based population-carriage alerts are used for infection-control decisions.

List of abbreviations

AMR, antimicrobial resistance; DiD, difference-in-differences; MC, Monte Carlo; pp, percentage points; R, resistant carriage; S, sensitive carriage; SR, simultaneous sensitive and resistant carriage. Cs and Cr are figure display labels for S and R, respectively.

Declarations

Ethics approval and consent to participate

Not applicable. This study used entirely synthetic observations and did not recruit participants or analyse human participant records or biological specimens.

Consent for publication

Not applicable.

Availability of data and materials

All observations in this study were simulated. Additional file 1 provides complete methods, equations, supporting results and figure legends. Additional file 2 contains aggregate scenario results, prediction-

library definitions, scenario partitions, fixed thresholds, fitting diagnostics and figure source data. No external participant-level dataset was analysed. No DHS records or other participant microdata were used.

Code availability

The simulation and figure-generation code is available from the corresponding author, Chaokun Hong (chaokun.hong@ndm.ox.ac.uk), on reasonable request.

Competing interests

The author declares no competing interests.

Funding

No funding was received for this research.

Authors' contributions

C.H. conceived and designed the study, developed the simulation and analysis framework, performed the analyses, prepared the figures, drafted and revised the manuscript, and approved the final version.

Acknowledgements

The author thanks the Demographic and Health Surveys Program and the participating national survey teams; the Climate Hazards Center for CHIRPS; the Copernicus Climate Change Service and the European Centre for Medium-Range Weather Forecasts for AgERA5 and GloFAS; the Global Flood Database team; the Uppsala Conflict Data Program; the United Nations High Commissioner for Refugees; and Our World in Data, the United Nations Inter-agency Group for Child Mortality Estimation and Gapminder for maintaining and making their data resources and documentation available. Those resources informed the broader research context; none supplied participant observations or empirical calibration data for the present study, which used exclusively synthetic data.

Clinical trial number: not applicable.

Tables

Table 1. Simulation design and principal assumptions

Component	Specification
Population target	Mean exposed-site resistant-carriage change versus unshocked counterfactual, days 0–179
Baseline carriage	S=0.56; R=0.14 in both structures
Joint carriage	A permits co-carriage; B has mutually exclusive S and R states
Sampling windows	Baseline −90 to −1; early 0–89; late 90–179
Cluster size and design effects	20 participants; observed-R design effect 1, 2 or 4
Diagnostic performance	S: sensitivity 0.90, specificity 0.95; R: sensitivity 0.90, specificity 0.97
Prediction library	742 trajectories; two baseline weights; five participation changes; 7420 rows
Budget ceiling	32000 synthetic cost units; survey unit cost 1, 3 or 8
Scenario partitions	Calibration 17; selection 29; evaluation 44
Replicates per scenario	4000; 1000; 3000, respectively
Additional fitting experiment	44 scenarios × 3 S-participation weights × 1000 replicates; fixed library and thresholds
Empirical fitting	None; baseline equilibrium matching, target matching and synthetic library fitting only
Evaluation summaries	Maximum false alarm across 14 nulls; mean power across 18 increases of +1 to +5 pp

Full continuous parameter ranges, generation seeds, realised costs, larger positive scenarios and supporting checks are provided in Additional files 1–2. Calibration branches and budgets are a separate supporting analysis and do not change the fixed-budget rule comparison.

Table 2. Participation width and threshold-calibration coverage

Calibration policy	Allowed weight change	Dynamic false alarm / power, %	No-projection false alarm / power, %
Within stated range	0	30.37 / 24.28	18.47 / 24.95
Within stated range	±0.15	10.50 / 25.11	7.93 / 28.26
Within stated range	±0.30	3.03 / 19.24	2.97 / 25.42
All original nulls	0	6.63 / 11.47	6.50 / 17.97
All original nulls	±0.15	3.87 / 16.12	3.67 / 22.76
All original nulls	±0.30	3.03 / 19.24	2.97 / 25.42

False alarm denotes the maximum over the same 14 null evaluation scenarios; power denotes the equal-weight mean over the same 18 increases of +1 to +5 pp. Each scenario used 3,000 replicates. The narrow within-range calibrations contain 12 nulls without differential participation, while the evaluation includes changes of −0.25 and +0.20. All-null calibration contains 17 nulls at every width. Uncertainty and scenario-specific values are in Additional file 2.

Table 3. Sensitivity to participation depending on sensitive carriage

S weight	Rule	Max false alarm (upper), %	Mean power, %
0.8	Early dynamic	3.8 (5.73)	19.04
0.8	Early no projection	11.1 (14.03)	19.73
0.8	Calibrated direct	9.5 (12.26)	15.55
1.0	Early dynamic	4.1 (6.09)	19.29
1.0	Early no projection	4.0 (5.97)	25.79
1.0	Calibrated direct	1.6 (2.99)	11.02
1.2	Early dynamic	77.2 (80.67)	12.23
1.2	Early no projection	76.9 (80.39)	26.16
1.2	Calibrated direct	1.9 (3.38)	8.27

All rules retain the full ±0.30 R-participation library and their original all-null thresholds. S weight applies only at the exposed site after baseline. False alarm is the maximum over 14 zero-target scenarios; the parenthesis is its simultaneous 95% Monte Carlo upper bound. Power is the equal-weight mean over 18 increases of +1 to +5 pp. Each scenario and S weight uses 1,000 new replicates. The S weight 1.0 reproduces the original observation formula with new random draws. Weights 0.8 and 1.2 are unrepresented sensitivity conditions, not empirical estimates. Full values and intervals are in Additional file 2.

Figure legends

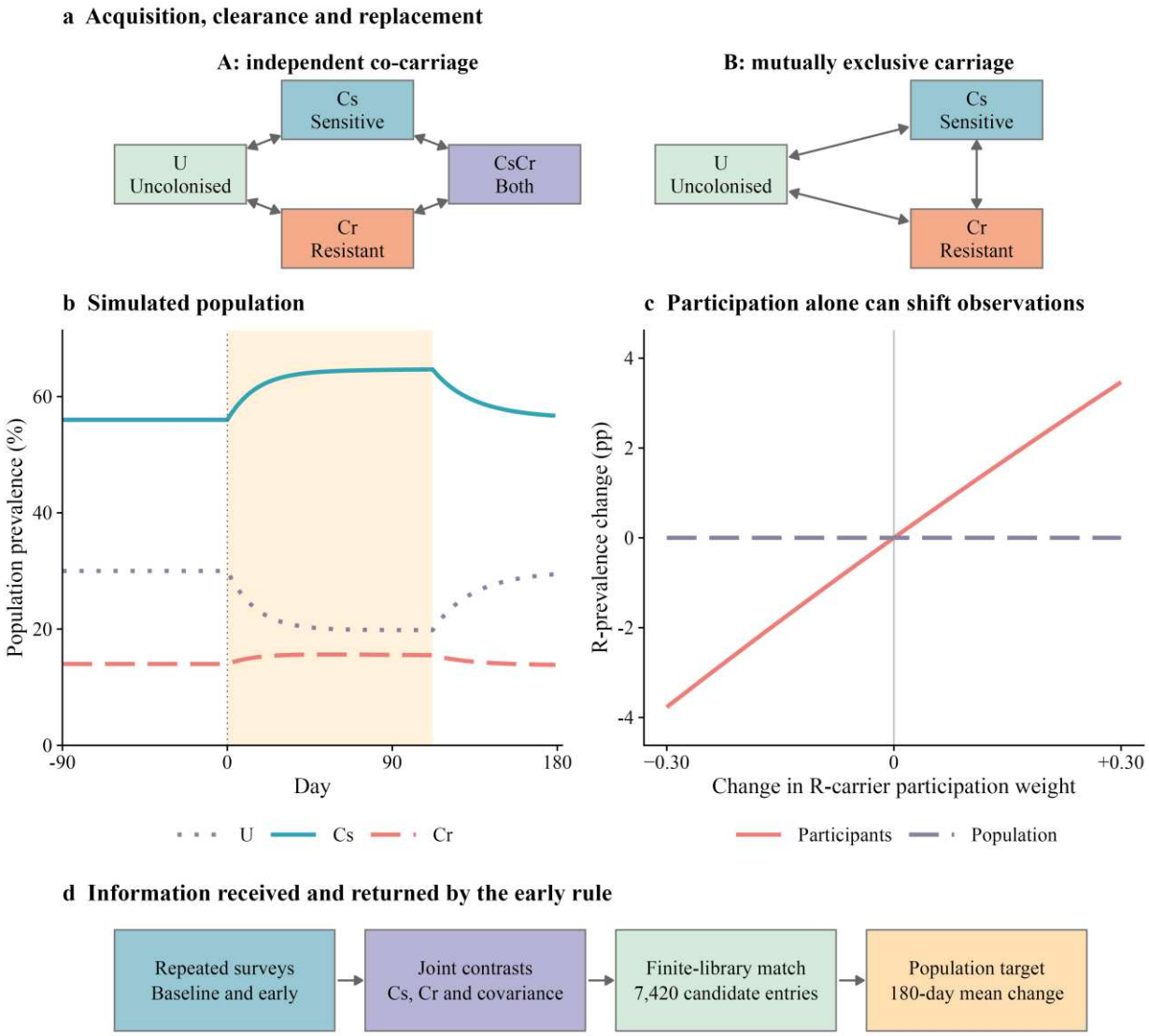


Figure 1. Carriage dynamics and the observation process underlying early surveillance. (a) Existing carriage structures. U denotes uncolonised, Cs sensitive carriage, Cr resistant carriage and CsCr co-carriage. Structure A permits independent acquisition and clearance of both types; structure B permits mutually exclusive carriage and replacement. Cs and Cr correspond to S and R in the equations; R does not denote recovery. (b) Daily exposed-site population prevalences in one existing synthetic scenario, starting from equilibrium. Shading marks the finite biological perturbation. These are generating trajectories, not daily observations supplied to the rule. (c) A separate calculation isolates participation: population resistant carriage remains 14%, baseline relative R-carrier participation weight is 1, and that weight changes by the amount shown. The non-carrier weight remains 1. The participant change follows Rw/(1−R+Rw); the population change remains zero. The R relationship is identical in both structures. Full sensitive/resistant calculations for both baseline weights appear in Additional file 1, Fig. S1. (d) The early rule receives two assay-corrected survey contrasts and their estimated covariance, matches them to a finite prediction library, and assigns the selected 180-day population target. All panels are synthetic or deterministic; no empirical time-series fit is shown.

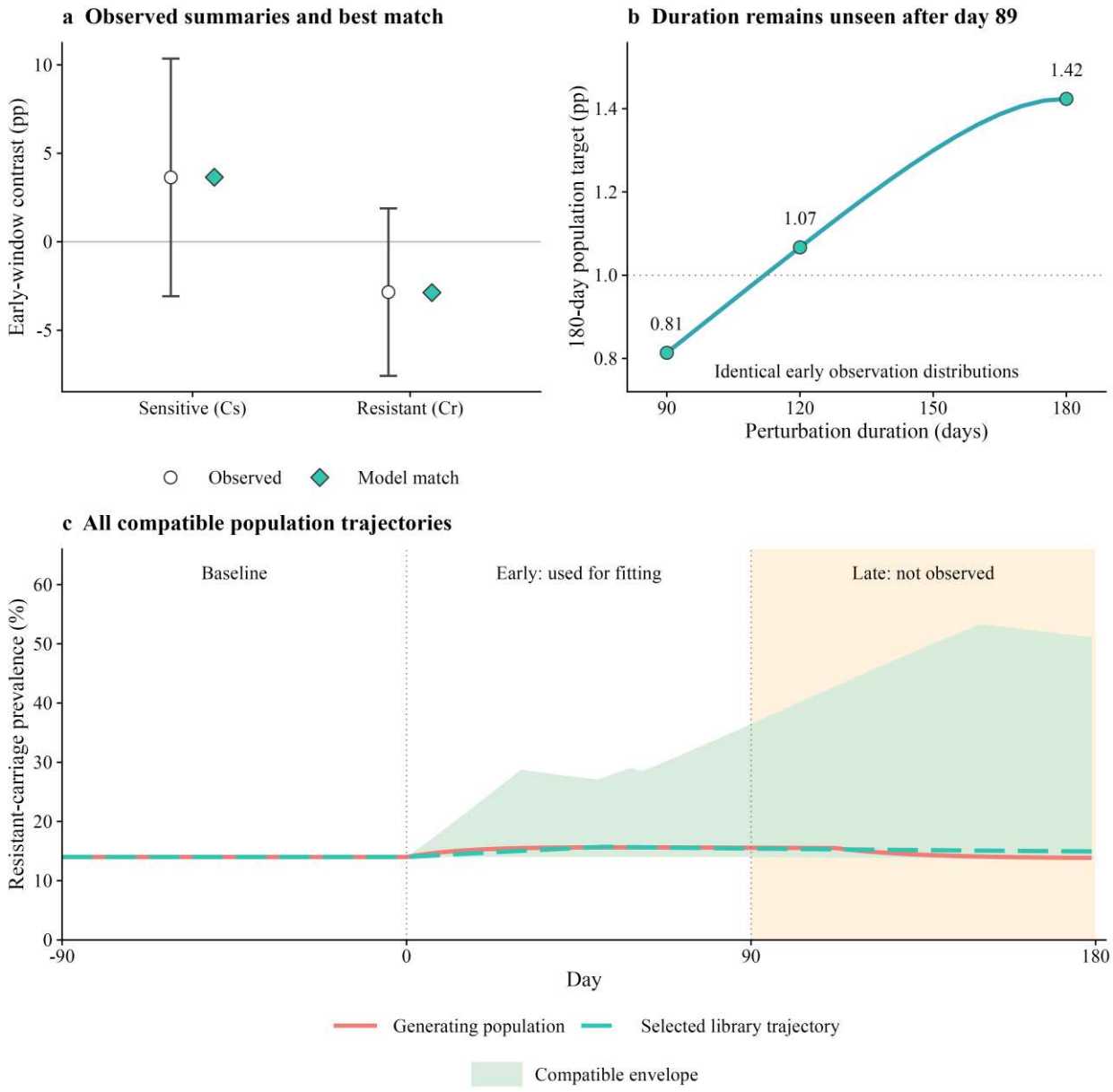


Figure 2. Early fit leaves the later population target unresolved. (a) Sensitive and resistant early survey contrasts from one synthetic replicate, with marginal nominal 95% t intervals, and predictions from its selected library entry. Joint matching uses the estimated covariance. (b) Holding the example's generating parameters fixed except perturbation duration, 19 durations from 90 to 180 days in five-day increments produce identical early observation distributions but different 180-day targets. The line connects computed deterministic values; markers identify 90, 120 and 180 days. The dotted horizontal line marks +1 pp. These noise-free comparisons are distinct from the noisy realisation in a. (c) Generating and selected resistant-carriage trajectories, with the pointwise minimum–maximum envelope of all 428 compatible trajectory templates, corresponding to 2,738 catalogue entries. Compatibility uses squared distance within 5.991 of the minimum. The envelope is not a confidence or credible interval and need not be attained by one template throughout time. The shaded late window is unobserved by the early rule. The illustrative scenario was selected by a fixed ID rule before plotting: first structure-B common-perturbation scenario with target +1 to +5 pp, S weight 1, replicate 1. No compatible template was omitted. This example and duration diagnostic establish information limits within the declared synthetic setting, not representative forecasting performance.

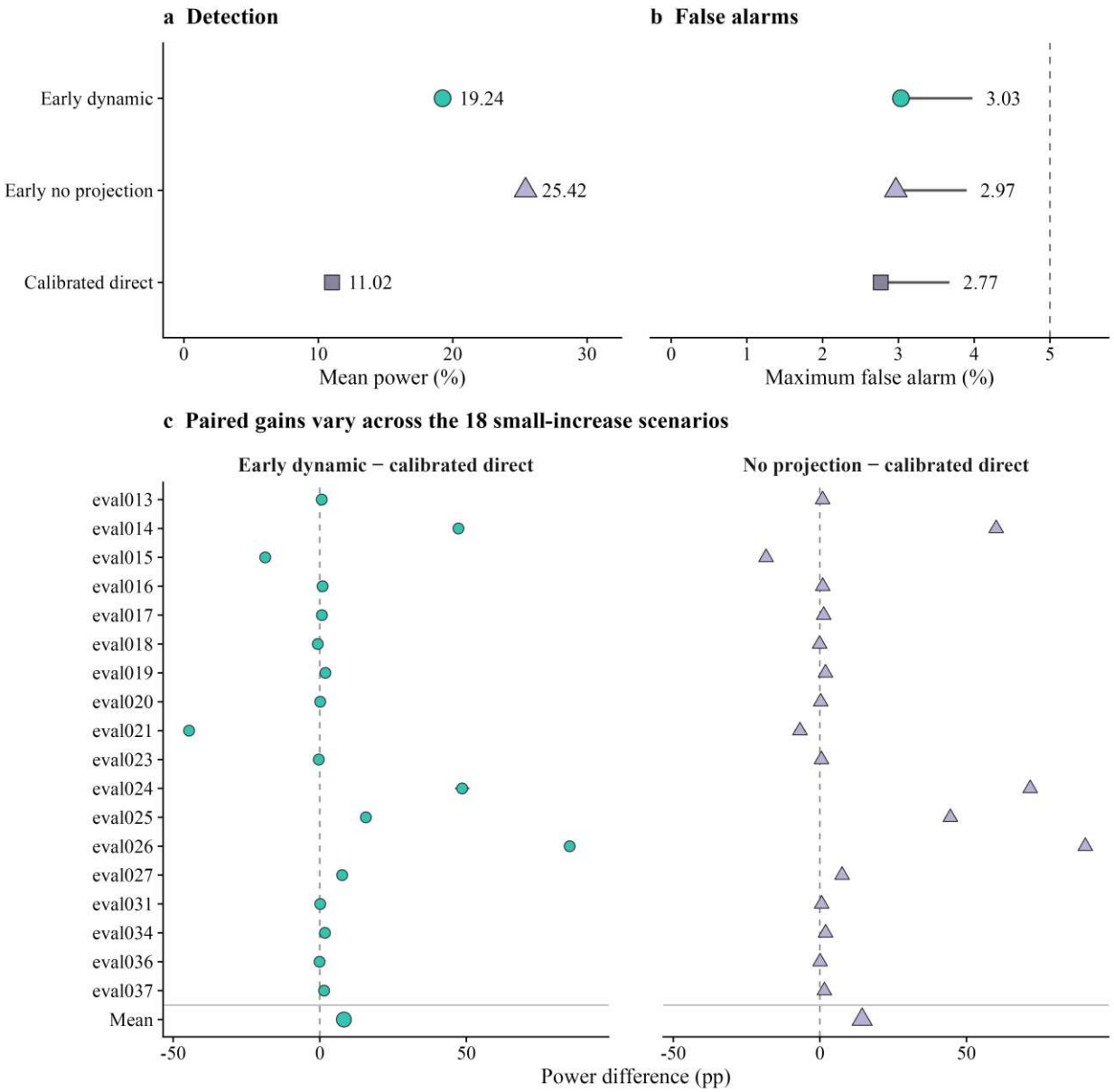


Figure 3. Calibrated rule performance and paired differences across scenarios. The early dynamic, early no-projection and direct two-window rules use the same null-calibration principle and a common budget ceiling, under the original R-only participation mechanism. (a) Equal-weight mean power over 18 population increases of +1 to +5 pp, with 95% Monte Carlo intervals. (b) Maximum false-alarm probability over 14 nulls; segments extend to simultaneous 95% Monte Carlo upper bounds within each rule. The dashed line marks 5%. Each scenario uses 3,000 replicates. Early rules collect baseline and early-window observations; the direct rule also observes the late window and allocates its budget across both follow-up windows. (c) Paired power differences from calibrated direct for every small-increase scenario, followed by their equal-weight mean. Segments are pointwise 95% Monte Carlo intervals retaining shared-sample dependence; the mean uses the variance of the equally weighted paired scenario differences. Some intervals are shorter than their symbols. Zero marks no difference. Negative values and scenario heterogeneity are retained. These intervals describe simulation precision for the fixed scenarios, not uncertainty about transfer to a real population. Nominal and clinical proxy references remain in Additional file 1, Fig. S5.

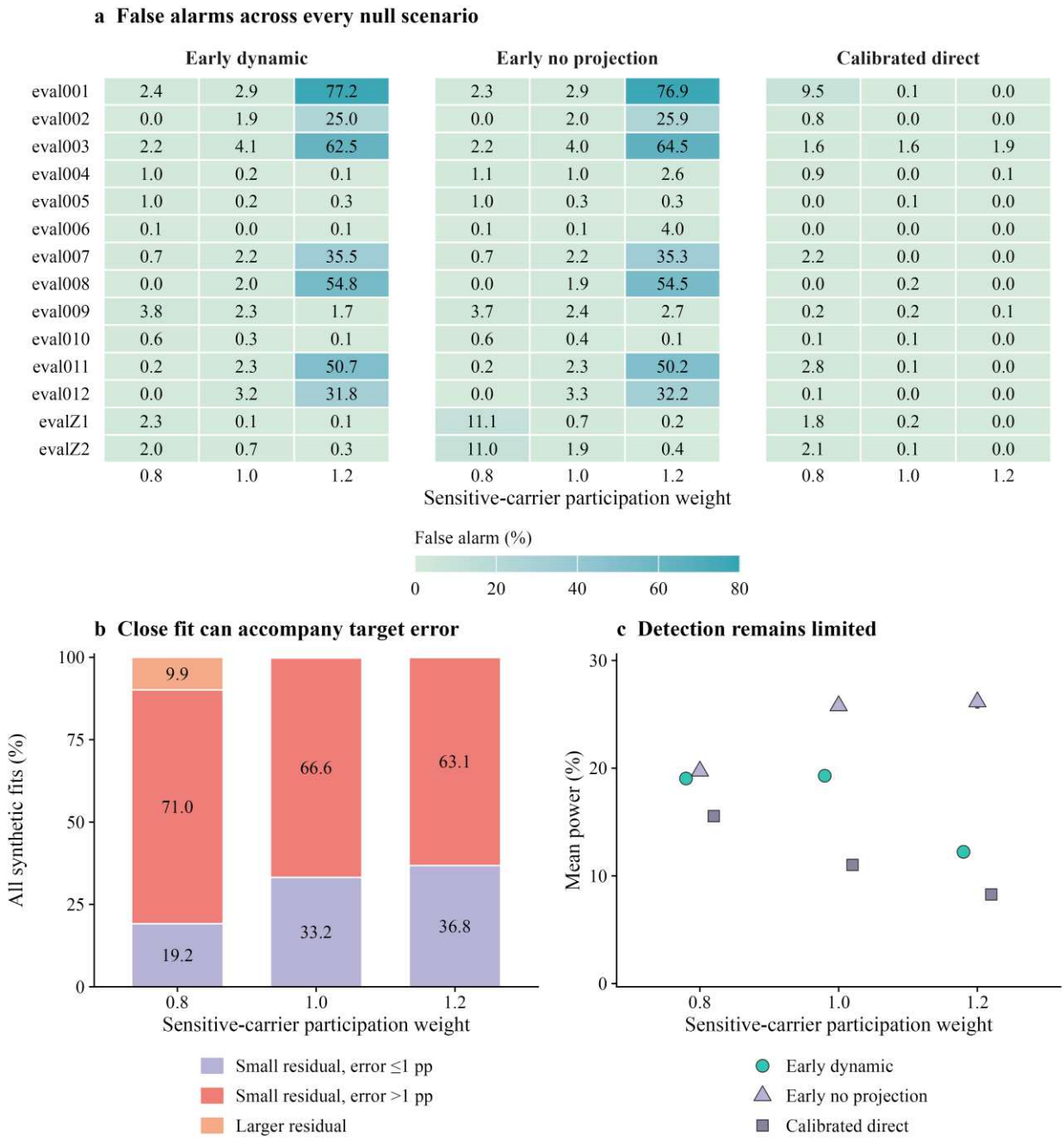


Figure 4. Selection mismatch defeats calibration despite close observed-summary fit. Sensitive-carrier participation weights vary only at the exposed site after baseline; biology, R-selection settings, the library and thresholds remain fixed. (a) False-alarm percentages for all 14 null scenarios, three weights and three rules, with 1,000 replicates per cell. Cells use a common 0–80% colour scale without clustering; their values are Monte Carlo point estimates. Maximum rates and simultaneous upper bounds are reported in Table 3. (b) Composition of all 44,000 fits at each weight, covering all 44 scenarios including larger increases. A small residual means minimum squared distance ≤5.991; target error is the absolute difference between fitted and generating 180-day population change. The categories partition all fits. Labels show percentages where space permits; the complete values are in Additional file 2. These are descriptive proportions, and the residual criterion is not a calibrated post-fit goodness-of-fit test. (c) Equal-weight mean power across 18 increases of +1 to +5 pp, with 95% Monte Carlo intervals, sometimes shorter than symbols. Weights are discrete sensitivity conditions; no interpolated response is assumed. Complete target-recovery and compatibility-width diagnostics appear in Additional file 1, Fig. S4.

References

1. World Health Organization. GLASS manual for antimicrobial resistance surveillance in common bacteria causing human infection. World Health Organization; 2023. https://www.who.int/publications/i/item/9789240076600

2. Rempel O, Pitout JD, Laupland KB. Antimicrobial Resistance Surveillance Systems: Are Potential Biases Taken into Account? Canadian Journal of Infectious Diseases and Medical Microbiology. 2011;22(4):e24–e28. https://doi.org/10.1155/2011/276017

3. Lim C, Hantrakun V, Teerawattanasook N, Srisamang P, Teparrukkul P, Sumpradit N, et al. Impact of low blood culture usage on rates of antimicrobial resistance. Journal of Infection. 2021;82(3):355–362. https://doi.org/10.1016/j.jinf.2020.10.040

4. Sinto R, Lie KC, Setiati S, Suwarto S, Nelwan EJ, Djumaryo DH, et al. Blood culture utilization and epidemiology of antimicrobial-resistant bloodstream infections before and during the COVID-19 pandemic in the Indonesian national referral hospital. Antimicrob Resist Infect Control. 2022;11:73. https://doi.org/10.1186/s13756-022-01114-x.

5. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592. https://doi.org/10.1093/biomet/63.3.581

6. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. Epidemiology. 2004;15(5):615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43

7. Manski CF. Anatomy of the Selection Problem. Journal of Human Resources. 1989;24(3):343–360. https://doi.org/10.2307/145818

8. Hicks AL, Kissler SM, Mortimer TD, Ma KC, Taiaroa G, Ashcroft M, et al. Targeted surveillance strategies for efficient detection of novel antibiotic resistance variants. eLife. 2020;9:e56367. https://doi.org/10.7554/eLife.56367

9. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279–4292. https://doi.org/10.1002/sim.2673

10. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074–2102. https://doi.org/10.1002/sim.8086

11. Abadie A. Semiparametric Difference-in-Differences Estimators. The Review of Economic Studies. 2005;72(1):1–19. https://doi.org/10.1111/0034-6527.00321

12. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE. 2012;7(12):e52078. https://doi.org/10.1371/journal.pone.0052078

13. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. American Journal of Epidemiology. 1978;107(1):71–76. https://doi.org/10.1093/oxfordjournals.aje.a112510

14. Welch BL. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. Biometrika. 1947;34(1-2):28–35. https://doi.org/10.1093/biomet/34.1-2.28

15. Mahalanobis PC. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India. 1936;2(1):49–55. https://hdl.handle.net/10263/6765

16. Glasserman P, Yao DD. Some Guidelines and Guarantees for Common Random Numbers. Management Science. 1992;38(6):884–908. https://doi.org/10.1287/mnsc.38.6.884

17. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika. 1934;26(4):404–413. https://doi.org/10.1093/biomet/26.4.404

18. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6(2):65–70. https://www.jstor.org/stable/4615733

19. R Core Team. R: A language and environment for statistical computing [software]. Vienna, Austria: R Foundation for Statistical Computing; 2025. https://www.R-project.org/

20. Soetaert K, Petzoldt T, Setzer RW. Solving Differential Equations in R: Package deSolve. Journal of Statistical Software. 2010;33(9):1–25. https://doi.org/10.18637/jss.v033.i09

21. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research. 2013;22(3):278–295. https://doi.org/10.1177/0962280210395740

22. Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association. 1993;88(421):125–134. https://doi.org/10.1080/01621459.1993.10594302

23. Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. The Lancet. 2016;387(10014):176–187. https://doi.org/10.1016/S0140-6736(15)00473-0

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.docx
- TableS2.docx
- FigureS1.tiff
- TableS3.docx
- FigureS2.tiff
- TableS4.docx
- FigureS3.tiff
- TableS5.docx
- FigureS4.tiff
- TableS6.docx
- FigureS5.tiff
- TableS7.docx
- ARICAdditionalfile1.docx