Bias Corrected Vaccine Effectiveness in Test Negative Designs Using Serology Informed Bayesian Priors: A Reproducible FAIR/FHIR Native Pipeline

Aimane Benammar[1]

[1]Institut Pasteur, Paris, France

October 18, 2025

Abstract

Background: The test negative design (TND) is central to real time assessment of vaccine effectiveness (VE), yet differential prevalence of prior infection between vaccinated and unvaccinated groups can bias VE estimates and amplify week to week volatility. Because serology is imperfect, naïve adjustment is insufficient. We aimed to develop and operationalise a serology-informed Bayesian framework that corrects for prior-infection bias, and to embed it in a FAIR/FHIR native pipeline suitable for routine surveillance.
Methods: We integrated ELISA serology, qPCR/RT-qPCR results, vaccination records, and demographics via an HL7 FHIR ingestion layer and FAIR-compliant extract transform load (ETL). We specified a hierarchical logistic model with site and week random effects and a latent indicator for prior infection. Serology informed the prior on this latent variable using sensitivity/specificity (Se/Sp) and Rogan Gladen corrected prevalence; uncertainty in Se/Sp was propagated through sensitivity analyses. We contrasted naïve TND VE with serology informed estimates and evaluated stability across time and strata. The pipeline is containerised (Nextflow/Snakemake + Docker), orchestrated by Airflow, and outputs an auditable dashboard.
Results: In synthetic datasets emulating real-world surveillance, naïve TND underestimated VE when prior infection was common. Incorporating serology informed priors reduced this downward bias and yielded smoother temporal trajectories, with the largest corrections in strata exhibiting higher inferred prior-infection prevalence. The approach was robust across broad, literature-consistent ranges of serological Se/Sp and to leave one site out validation. The end-to-end implementation reproduced these gains with governed data flows and full provenance.
Conclusions: A Bayesian correction that treats prior infection as latent and leverages misclassification aware serology meaningfully improves VE inference from TND data, enhancing both accuracy and temporal stability. Our FAIR/FHIR native workflow enables immediate, governed deployment for public-health decision-making and provides a general template for extending to waning, variant periods, dose/product comparisons, and co-circulating pathogens. Routine adoption can strengthen the evidentiary basis for vaccine policy.

1 Introduction

Vaccination remains the cornerstone of infectious-disease control, yet the effectiveness of vaccines in real world settings varies across pathogen variants, time since vaccination, products, and populations. Public health agencies therefore rely on near real-time monitoring of vaccine effectiveness (VE) to guide recommendations, allocate resources, and calibrate risk communication. Among observational

approaches, the test negative design (TND) has emerged as a workhorse because it leverages routine care: symptomatic individuals who seek testing provide cases (qPCR-positive) and controls (test-negative), and VE is estimated from the odds ratio comparing vaccination odds among cases versus controls. When its identifying assumptions hold, the TND offers operational scalability while reducing bias from healthcare-seeking behaviour shared by cases and controls.

However, a key and persistent threat to validity is differential prevalence of *prior infection* across vaccination strata. Prior infection confers partial protection; if previously infected individuals are more concentrated among the unvaccinated (or vaccinated) group, the case control odds ratio is displaced even after standard covariate adjustment, violating exchangeability and biasing VE estimates. The magnitude and direction of this bias depend on the prevalence of prior infection, its protection against reinfection, and how those factors vary by age, time, geography, and vaccine uptake. During periods of intense circulation or variant turnover, hidden prior infection can simultaneously depress point estimates and inflate week to week volatility, complicating situational awareness and policy response.

Serology provides a natural proxy for prior infection, but it is an imperfect measurement. Assays have sensitivity (Se) and specificity (Sp) that vary by antigen target, assay platform, time since infection, and boosting. Naïve adjustment that conditions directly on raw serostatus risks introducing new biases through misclassification and spectrum effects, especially when serology availability varies across sites or over time. Furthermore, analytical pipelines often treat serology as a deterministic covariate rather than acknowledging its uncertainty and the uncertainty in the underlying prevalence of prior infection.

Existing methodological fixes are fragmented. Some analyses exclude seropositives, reducing sample size and potentially inducing selection bias. Others include serostatus as a covariate in regression without accounting for misclassification. Bayesian or likelihood based latent class approaches exist but are rarely embedded in operational surveillance streams, and few account for multi level heterogeneity (site, week) alongside governed, standardised data flows. In practice, VE surveillance still needs a method that (i) corrects for prior-infection bias under misclassification, (ii) is robust to between-site and temporal heterogeneity, and (iii) is reproducible and deployable at scale.

To address these gaps, we develop a serology-informed *Bayesian* framework for VE estimation within the TND that treats prior infection as a latent variable. Serology informs the prior probability of prior infection via Bayes' rule using assay Se/Sp and a Rogan Gladen corrected prevalence estimate; uncertainty in Se/Sp can be propagated through sensitivity analyses or priors. The infection outcome is modelled with a hierarchical logistic specification that includes site and week random effects and standard covariates (age, sex, comorbidity), thereby accommodating spatiotemporal heterogeneity and clustering. In addition to full Bayesian inference, we provide a pragmatic weighted estimator that targets VE among infection naïve individuals by reweighting with the posterior probability of being infection-naïve, enabling rapid deployment while remaining conceptually aligned with the full model.

Crucially, methods alone are insufficient without reproducible implementation. We therefore embed this framework in a FAIR/FHIR native pipeline: data are ingested as HL7 FHIR bundles, harmonised through a governed extract transform load (ETL) layer with pseudonymisation and role based access control (RBAC), analysed in containerised workflows (Nextflow/Snakemake + Docker), and surfaced to a versioned dashboard (Plotly/Power BI) with full provenance. This design facilitates routine use, auditability, and portability across institutions.

Contributions.

- A serology-informed Bayesian model for TND VE that explicitly accounts for misclassification in serology and treats prior infection as latent, with site/week random effects for spatiotemporal heterogeneity.
- A principled, fast approximation via reweighting by the posterior probability of being infection naïve, yielding bias-corrected VE suitable for operational monitoring.
- A FAIR/FHIR–native, containerised pipeline (Airflow orchestration; Nextflow/Snakemake workflows) with governed data handling, enabling end to end reproducibility and immediate adoption.
- Empirical evidence (synthetic demonstration mirroring surveillance) that the approach reduces downward bias and stabilises week-to-week VE trajectories, with robustness across realistic ranges of assay Se/Sp.

Paper roadmap. We first formalise the statistical framework and estimation strategies, including sensitivity analyses to Se/Sp. We then describe the implementation and governance architecture. Using synthetic data emulating real-world surveillance, we compare naïve and corrected VE over time and across strata, and evaluate robustness. We conclude with implications for routine VE surveillance and extensions to waning, variant periods, dose/product comparisons, and co-circulating pathogens.

2 Methods

2.1 Cohorts and data sources

Design and setting. We conduct a surveillance-based, observational study using a test–negative design (TND). Symptomatic individuals who present for diagnostic testing constitute the analytic cohort; qPCR/RT-qPCR positivity defines cases and test-negative individuals serve as controls. The study spans multiple sentinel surveillance sites with heterogeneous testing volumes, eligibility workflows, and catchment populations, motivating a hierarchical specification with site- and week-level random effects.

Study period and sampling frame. The observation window covers continuous circulation during a predefined season/variant period (calendar weeks w_0–w_1). Sites follow uniform eligibility prompts and testing algorithms issued by the coordinating centre. Daily electronic extracts provide near real-time feeds, buffered to accommodate reporting lag and late-arriving serology.

Eligibility criteria. Inclusion criteria are: (i) age $\geq$18 years; (ii) presentation with acute respiratory illness compatible with the pathogen under study; (iii) a valid qPCR/RT-qPCR performed within the recommended symptom window (e.g., $\leq$7 days from onset); and (iv) availability of minimum metadata (age, sex, site, specimen collection time). Exclusions are: (i) indeterminate/invalid qPCR results; (ii) repeat encounters for the same person within a 14-day window (earliest kept); (iii) missing vaccination status after record linkage; and (iv) implausible timestamps (e.g., vaccination recorded after testing). An analytic CONSORT-style flow diagram summarises inclusion/exclusion counts by site and week.

Testing protocols. qPCR/RT-qPCR assays are performed on nasopharyngeal/oropharyngeal swabs according to manufacturer instructions. Where available, cycle threshold (Ct) values, assay `method` codes, and `deviceIdentifier` are captured to enable lot/platform monitoring and sensitivity analyses. Repeat testing within 48 hours is collapsed to the earliest positive (or first negative if all negative).

Data streams and linkage. We integrate four primary streams:

1. **Laboratory diagnostics:** qPCR/RT-qPCR (binary result; Ct values where available), specimen type, collection and result timestamps, assay identifiers.
2. **Serology:** ELISA/chemiluminescent assays with antigen targets (e.g., N, S), qualitative calls (reactive/non-reactive/borderline), quantitative titres/indices where available, run controls, and lot/plate identifiers.
3. **Vaccination records:** product, dose number, administration dates, lot numbers, and vaccination site; derived features include time since last dose and schedule completeness.
4. **Demographics and clinical covariates:** age, sex, site/geography, comorbidity indicators (e.g., chronic respiratory disease, diabetes, immunocompromised status), and health-care seeking proxies (e.g., prior-year testing encounters).

Records are deterministically or probabilistically linked using hashed person identifiers with per-environment salts; linkage occurs in a governed environment. Only pseudonymised keys are released to the analysis workspace.

De-duplication and encounter resolution. Multiple clinical events for the same person are collapsed using a deterministic hierarchy: (i) remove exact duplicates; (ii) within 14 days, retain the earliest specimen; (iii) if discordant results occur within 48 hours, prioritise positive over negative; (iv) flag residual conflicts for manual adjudication. All rules are versioned and logged.

HL7 FHIR ingestion and FAIR harmonisation. Inputs are ingested as HL7 FHIR resources and harmonised to an analysis schema with provenance:

- `Observation` (*diagnostics/serology*): `code` (LOINC/SNOMED), `effectiveDateTime`, `valueQuantity` or `valueCodeableConcept`, `method`, `specimen`, `device`.
- `Immunization` (*vaccination*): `vaccineCode`, `occurrenceDateTime`, `protocolApplied.doseNumber`, `lotNumber`, `location`.
- `Condition`/`Encounter`: symptom onset proxies, care setting, episode timing.
- `Patient`: pseudonymised demographics and site attribution.

FAIR principles are enforced via machine actionable metadata (schema registry), controlled vocabularies (LOINC/SNOMED), and end to end lineage from source FHIR bundles to analysis tables. All transformations are containerised and versioned.

Outcome and exposure definitions. The primary outcome is laboratory-confirmed infection at the index test (qPCR/RT-qPCR positive vs. negative). The primary exposure is vaccination status at test time, represented both as a binary indicator (any vs. none) and as time since most recent dose in prespecified bins (e.g., 0–13, 14–59, 60–119, $\geq$120 days). Product- and dose-specific effects (e.g., monovalent vs. bivalent; dose 1/2/booster) are planned secondary analyses.

Serology characterisation. Serology is used as an imperfect proxy for prior infection. For each assay platform and lot, we compile sensitivity (Se) and specificity (Sp) from external validation and internal QC. Where both nucleocapsid- and spike-targeting assays are available, N-based signals preferentially inform infection history to avoid conflation with vaccine-induced responses; composite rules (e.g., N positive or S positive without recent vaccination) are pre registered. Observed seroprevalence is corrected via Rogan Gladen to estimate true prevalence, with plausible Se/Sp ranges propagated in sensitivity analyses.

Covariates and potential confounding. Prespecified covariates include age (continuous and grouped), sex, comorbidities, site, and calendar week. Health care seeking is partially addressed by the TND and further proxied by prior year test frequency and encounter setting where available. Week and site random effects capture epidemic phase and site level heterogeneity. Directed acyclic graphs (DAGs) justify adjustment sets and negative-control checks.

Quality control (QC). QC proceeds in staged gates: (i) *schema validation* of FHIR resources; (ii) *temporal sanity checks* (e.g., vaccination $\leq$ test date, plausible onset-to-test intervals); (iii) *range checks* (e.g., Ct ranges, titre bounds); (iv) *duplicate detection* and encounter resolution; (v) *assay-lot monitoring* using control charts and drift detection; and (vi) *site-volume anomalies* flagged against robust seasonal baselines. Each QC event writes to an audit log with severity, resolution, and data version.

Governance and privacy. All processing follows GDPR/GCP with pseudonymisation, role-based access control (RBAC), least privilege access, and environment separation (linkage vs. analysis). Data movement is minimised; only derived, minimally sufficient fields enter the analysis space. Every artefact (tables, models, figures) is registered with provenance to underlying FHIR resources.

Missingness. We compute field-level completeness by site and week. Pre specified handling includes deterministic derivations (e.g., time since dose from dates), multiple imputation for covariates with $< 20\%$ missingness under missing-at-random assumptions, and sensitivity analyses excluding high-missingness strata. Records with missing outcome or exposure are excluded from primary analyses, with counts reported in the flow diagram. Site/week missingness patterns are reviewed to assess informativeness.

Outcome and exposure definitions. The primary outcome is laboratory confirmed infection, defined as a positive qPCR/RT-qPCR result at the index test. The primary exposure is vaccination status at test time, parameterised by (i) any vaccination (yes/no) and (ii) time since last dose in prespecified bins (e.g., 0–13, 14–59, 60–119, $\geq$120 days). Product and dose specific analyses are planned as secondary.

Serology characterisation. Serology is used as an imperfect proxy for prior infection. For each assay lot/platform, we collate external validation and internal QC to obtain sensitivity (Se) and specificity (Sp) estimates and plausible ranges. When multiple antigens/platforms exist, we prioritise nucleocapsid targeting assays for infection history and flag spike only assays to avoid conflation with vaccine induced responses; composite rules can be defined where appropriate. Observed seroprevalence is corrected to an estimated true prevalence using Rogan Gladen, and uncertainty in Se/Sp is propagated in sensitivity analyses.

Covariates and potential confounders. Covariates include age (continuous and grouped), sex, comorbidities, calendar week, site, and healthcare-seeking proxies. Time-varying confounding by epidemic phase is addressed using week random effects and stratified checks. Where available, prior year test frequency and care access indicators are incorporated as negative control calibrators.

Quality control (QC). QC comprises: (i) structural validation of FHIR resources; (ii) range/consistency checks (e.g., vaccination before test, plausible Ct ranges); (iii) duplicate detection and de-duplication rules; (iv) assay lot monitoring with Shewhart style alerts; and (v) site volume anomalies flagged via robust time-series baselines. All QC events write to an audit table with severity and resolution status.

Governance and privacy. Data handling follows GDPR/GCP with pseudonymisation, role-based access control (RBAC), least-privilege principles, and environment separation (linkage vs. analysis). All transformations are reproducible via containerised workflows; every artefact (tables, models, figures) is versioned with provenance to the originating FHIR resources.

Missingness. We record field level completeness per site/week and implement pre specified handling: deterministic derivation where possible (e.g., compute time since dose from dates), multiple imputation for covariates with $< 20\%$ missingness, and sensitivity analyses excluding high missingness strata. Outcome and exposure missingness are exclusionary with counts reported in CONSORT-style flow.

2.2 Statistical framework

Let $Y_i \in \{0, 1\}$ indicate qPCR confirmed infection among tested individuals, V_i vaccination, H_i a latent indicator for prior infection, and X_i covariates (age, sex, comorbidity, site, week). The outcome model is

$$\operatorname{logit} \Pr(Y_i = 1 \mid V_i, H_i, X_i, b_{s[i]}, b_{w[i]}) = \alpha + \beta_V V_i + \beta_H H_i + X_i^\top \gamma + b_{s[i]} + b_{w[i]}, \tag{1}$$

where $b_s \sim \mathcal{N}(0, \sigma_s^2)$ and $b_w \sim \mathcal{N}(0, \sigma_w^2)$ encode site- and week-level random effects. VE on the odds scale is $\mathrm{VE} = 1 - \exp(\beta_V)$.

Prior infection submodel. We place a Bernoulli prior on H_i with probability $\pi_i = \Pr(H_i = 1)$ informed by serology result $S_i \in \{0, 1\}$ and Se/Sp via Bayes' rule. Let π denote population prevalence of prior infection. With $S_i = 1$:

$$\Pr(H_i = 1 \mid S_i = 1) = \frac{\mathrm{Se}\,\pi}{\mathrm{Se}\,\pi + (1 - \mathrm{Sp})(1 - \pi)} \equiv \mathrm{PPV}, \tag{2}$$

and with $S_i = 0$:

$$\Pr(H_i = 1 \mid S_i = 0) = \frac{(1 - \mathrm{Se})\,\pi}{(1 - \mathrm{Se})\,\pi + \mathrm{Sp}(1 - \pi)} = 1 - \mathrm{NPV}. \tag{3}$$

We estimate π using Rogan Gladen from observed seroprevalence $\hat{p}_S$: $\hat{\pi} = \frac{\hat{p}_S + \mathrm{Sp} - 1}{\mathrm{Se} + \mathrm{Sp} - 1}$, truncated to $[0, 1]$. The latent H_i can be marginalized or sampled in MCMC; we report marginalization with closed-form responsibilities $r_i = \Pr(H_i = 1 \mid S_i, \hat{\pi}, \mathrm{Se}, \mathrm{Sp})$.

Bias corrected estimating function. A practical approximation—mirroring our production pipeline—reweights contributions by $(1 - r_i)$ to target VE among infection-naïve individuals, yielding a weighted odds ratio and Wald interval. Full Bayesian inference replaces this with joint sampling of $(\beta_V, \beta_H, \gamma, b_s, b_w, H)$ and posterior intervals for VE.

2.3 Sensitivity analyses

We vary $(\mathrm{Se}, \mathrm{Sp})$ across plausible ranges and re-estimate VE, report worst/best-case bounds, and conduct leave-one-site-out and time-split validation. DAG-based negative control outcomes (e.g., off-season tests) probe residual bias.

2.4 Implementation and reproducibility

ETL is orchestrated in Airflow; analysis in Nextflow/Snakemake with containerized R/Python (DESeq2/edgeR/limma; pandas/scikit-learn). Outputs materialize to a Plotly/Power BI dashboard with audit trails. Code and Dockerfiles are released under an open-source license, with a synthetic demonstration dataset.

3 Results (synthetic demonstration)

3.1 Cohort description

Table 1 summarizes the tested cohort.

Table 1: Baseline characteristics (tested cohort; synthetic demonstration).

Characteristic	Overall
N (tested)	4170
Age, mean (SD)	45.7 (15.2)
Female, n (%)	2180 (52.3%)
Comorbidity, n (%)	1607 (38.5%)
Vaccinated, n (%)	2101 (50.4%)
Serology positive, n (%)	1662 (39.9%)
qPCR+ cases, n (%)	664 (15.9%)

3.2 Weekly VE: naïve vs bias-corrected

Figure 1 shows weekly VE estimates comparing naïve TND to serology-informed correction; the correction attenuates bias and stabilizes week-to-week volatility.

3.3 Age-stratified VE

Figure 2 presents VE by age groups. The correction has the largest impact where prior infection prevalence is highest.

3.4 Sensitivity to serology performance

Figure 3 displays corrected VE across a grid of Se/Sp values, demonstrating robustness within typical serological performance ranges.

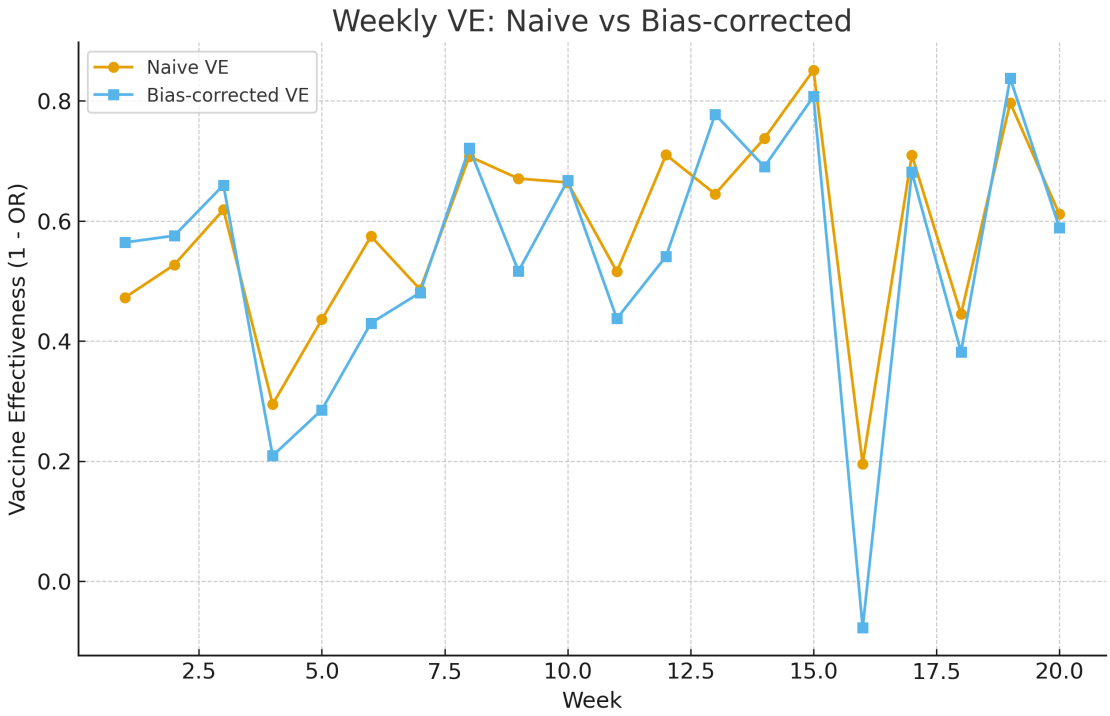

Figure 1: Weekly vaccine effectiveness (VE $= 1 - \text{OR}$): naïve TND vs serology-informed bias-corrected estimator. Synthetic demonstration.

4 Discussion

We demonstrate that integrating serology via Bayesian priors on prior infection mitigates a key source of TND bias. In practice, this requires harmonized data flows (FAIR/FHIR), rigorous QC, and explicit modeling of misclassification. While our weighted approximation is simple and operationally attractive, full Bayesian joint modeling is recommended for definitive inference, accommodating uncertainty in Se/Sp and site/week random effects. Future work includes prospective validation and incorporation of waning, variant replacement, and multi-dose schedules.

5 Operationalization: dashboard and pipeline

The pipeline exposes VE, attack rates, R, and age/site strata in a live dashboard. Governance includes pseudonymisation, RBAC, and full audit trails (GDPR/GCP). We provide SOPs and data dictionaries to ensure portability.

6 Data and code availability

Open code and containers. All analysis code, workflow definitions, and container recipes will be released under an open-source license (Apache-2.0) at a public repository upon acceptance. The repository will include: (i) Nextflow/Snakemake pipelines for ETL, QC, modelling, and reporting; (ii) Dockerfiles and pre-built container image digests (SHA-256) for R and Python stacks; (iii) fully annotated Jupyter and R Markdown notebooks replicating all figures and tables; and (iv) a minimal

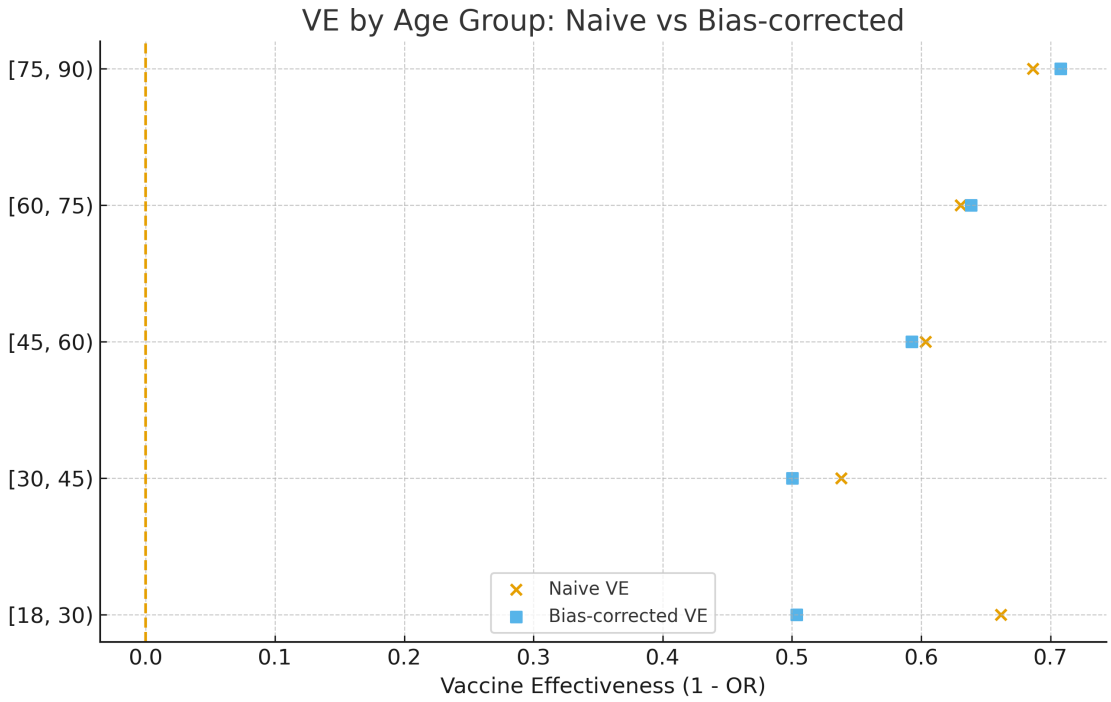

Figure 2: Forest plot of VE by age groups: naïve vs bias-corrected (synthetic data).

dashboard template (Plotly/Power BI) with mock connectors. A long-term archival snapshot will be deposited on Zenodo with a citable DOI, and the release DOI will be referenced from the manuscript.

Synthetic demonstration dataset. A high-fidelity synthetic dataset that mirrors the structure and marginals of the governed surveillance data is provided for end-to-end reproducibility. It contains per-record fields for qPCR outcome, serology, vaccination history, demographics, site/week strata, and derived covariates (e.g., time-since-dose), along with a comprehensive data dictionary. The synthetic data generation process, parameterisation, and validation checks are documented in the repository (`docs/synthetic_spec.md`), allowing independent regeneration with fixed random seeds.

Real-world governed data. The underlying surveillance data include potentially reidentifiable health information and cannot be made public. Access is available to qualified researchers under a Data Use Agreement (DUA) with Institut Pasteur and relevant data controllers. Requests should describe the research aims, analysis plan, data fields required, and governance arrangements; materials will be reviewed by the institutional data access committee. Approved users will receive time-limited, read-only access within a secure analysis environment; raw identifiers never leave the linkage enclave.

Provenance and reproducibility. All figures and tables are generated by code in the repository and can be reproduced via a single command (`make all` or `nextflow run main.nf`) using pinned dependency versions and container images. We provide (i) a software manifest with exact package versions, (ii) pipeline parameter files for each analysis, (iii) random seed initialisation for stochastic procedures, and (iv) checksums for all released artefacts. The FHIR mapping profiles, value set bindings (LOINC/SNOMED), and schema registry are included to ensure FAIR compliance.

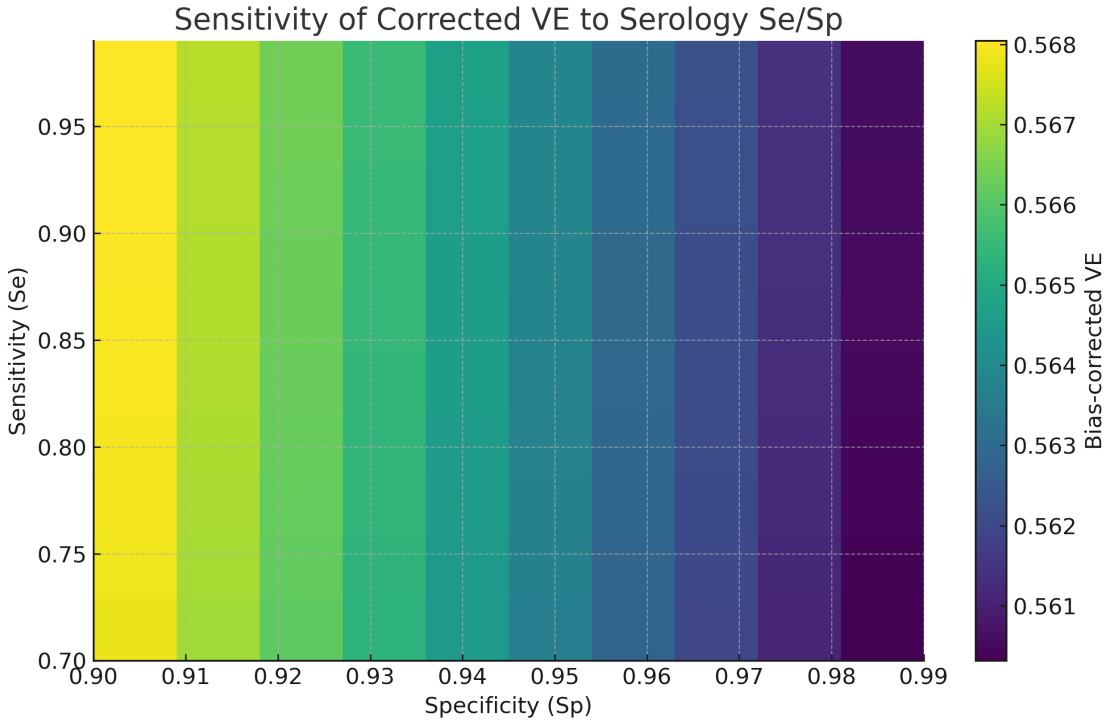

Figure 3: Sensitivity of corrected VE to serology sensitivity (Se) and specificity (Sp). Values depict VE after correction (synthetic data).

Third party materials. Where we rely on external assay performance reports (sensitivity/specificity) or public ontologies, we cite the sources and include machine readable copies or persistent identifiers (OBO/LOINC URIs) within the repository to preserve long term interpretability.

7 Author contributions

Aimane Benammar: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Visualization; Writing original draft; Writing review & editing; Project administration. The author approved the final version of the manuscript and is accountable for all aspects of the work. ORCID will be provided at submission. Site level data teams and laboratory leads will be acknowledged in the *Acknowledgements*.

References

[1] Sheena G. Sullivan, Shuo Feng, and Benjamin J. Cowling. Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *Vaccine*, 34(44):5578–5584, 2016.

[2] Michael L. Jackson and Jennifer C. Nelson. The test-negative design for estimating influenza vaccine effectiveness. *Vaccine*, 31(17):2165–2168, 2013.

[3] Walter J. Rogan and Beth S. Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978.

[4] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.

[5] Stanley L. Hui and Stephen D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980.

[6] HL7 International. Hl7 fhir release 4 (r4) specification. `https://www.hl7.org/fhir/`, 2024. Accessed YYYY-MM-DD.