University of Cape Town

Mathematical Modelling of Infectious Diseases

Rubella Vaccination in South Africa: Modelling CRS Impact, Timing, and Paradox Risk

October 19, 2025

Contents

Abstract

Background: Rubella remains endemic in South Africa, posing risk of congenital rubella syndrome (CRS). We assess whether introducing rubella-containing vaccine (RCV) with routine immunisation and a one-time catch-up reduces CRS over 10- and 25-year horizons while avoiding the "paradox" of increased infections among women of reproductive age. **Methods:** We build an age/sex-structured SEIR model with annual demographic events, routine vaccination at age 1, and a configurable catch-up. CRS is mapped from female infections (15–44y) using trimester-weighted risks and an age-specific pregnancy-prevalence proxy. Literature anchors transmissibility and reporting ($R_0 \in \{6, 7, 8\}$; high Vaccine Efficacy (VE)), with mean-level rescaling used only for validation. We validated magnitude and trend against surveillance and performed an *exploratory negative-binomial likelihood fit*; main policy conclusions use literature-anchored parameters, with NB results reported as sensitivity (Section 5.4). We explore routine vaccination (50–95% coverage), catch-up vaccination (0–95% coverage), and timing (immediate vs 1–3 year delays). **Results:** Higher routine plus high, wide catch-up (1–14y) gives the largest CRS reductions. With 90% catch-up, break-even routine (25 year averted >0 vs 80% routine/no catch-up) is ≈55–60% (R_0=6), ≈50% (R_0=7), and ≈80% (R_0=8). Low routine with modest catch-up risks the paradox. At routine 90%/catch-up 90%, immediate campaigns dominate at 10 year horizon and remain slightly better at 25 year horizon. **Conclusions:** Introduce RCV with ≥90% routine and a high-coverage 1–14y catch-up; in higher-transmission settings ($R_0 \geq 8$), avoid introduction without ≥80% routine or an adequately broad, well-covered catch-up.

1 Policy Question & Context

Question. Should South Africa introduce rubella-containing vaccine (RCV) now, and if so, with what combination of routine coverage and a one-time catch-up to minimise CRS over 10 and 25 years while avoiding the vaccination "paradox" [3, 5]? **Context.** Rubella's transmissibility typically lies in the mid–single digits ($R_0 \approx 5$–8) [1, 2, 5]. High routine coverage plus a broad catch-up collapses transmission and CRS; poorly covered roll-outs can shift average age at infection upward, raising infections in reproductive ages despite fewer total infections [3, 5]. We therefore focus on cumulative CRS (10/25y), cumulative infections in females 15–44 (paradox check), and timing trade-offs.

2 Literature Review

2.1 Biology and Transmission Dynamics

Rubella virus (genus *Rubivirus*) is a human-limited, droplet-transmitted infection with a short infectious period spanning roughly from just before rash onset to about a week afterward; subclinical infections are common, complicating surveillance and inference [5, 1]. In pre-vaccination settings, transmission is driven by high contact rates among children, with susceptibles replenished primarily via births. Maternal infection in early pregnancy can lead to fetal infection and congenital rubella syndrome (CRS), characterised by cardiac, ocular,

auditory and neurodevelopmental sequelae.

For policy modelling, the age structure of contacts and demography (birth and death rates) are central: reducing force of infection through vaccination suppresses transmission but can also increase the mean age at infection. If routine coverage is insufficient and historical susceptibility gaps persist, the average age shift can raise infections among women of reproductive age—the "vaccination paradox"—even as overall infections fall [3, 5]. This mechanism motivates explicit monitoring of infections in females aged 15–44 and of mean age at infection in that group.

2.2 Historical Burden and Epidemiology (Then vs Now)

Before widespread vaccination, rubella generated recurrent epidemics with occasional large CRS outbreaks when susceptible cohorts overlapped with transmission peaks [1, 2]. Countries introducing rubella-containing vaccine (RCV) with high routine coverage and broad catch-up campaigns rapidly reduced rubella and CRS, and many have achieved or neared elimination targets [5]. Conversely, where introduction occurred at low routine coverage or without adequate catch-up, models and experience have highlighted temporary increases in the proportion (and sometimes number) of infections in women of reproductive age, emphasising the need for careful rollout design [3, 5]. These observations underpin WHO's guidance to link introduction to programme capacity for high coverage and, where needed, broad catch-up to close historical immunity gaps.

2.3 Control: Treatment, Vaccination, Immunity

There is no antiviral therapy with population-level impact on transmission; control relies on vaccination, surveillance, and ensuring immunity among women of childbearing potential [5]. WHO recommends introducing RCV when programmes can achieve and sustain high routine coverage (typically $\geq 80{-}90\%$) and pairing introduction with a one-time, high-coverage catch-up across a wide age range (often 1–14 years) where susceptibility pockets are likely [5]. First-dose vaccine effectiveness against rubella infection is high (commonly $\geq 90{-}95\%$), and infection- or vaccine-derived immunity is generally long-lasting [5, 1].

For burden estimation, CRS risk is strongly time-dependent in pregnancy: it is highest with maternal infection in the first trimester and declines thereafter. In applied models, trimester-weighted risks provide a pragmatic approximation, mapping age-specific female infections to expected CRS cases when combined with an age profile of pregnancy prevalence [5]. This approach supports policy questions central to South Africa's introduction decision: (i) how routine coverage and the design of a catch-up campaign trade off in preventing CRS over 10- and 25-year horizons; and (ii) which combinations risk triggering the paradox in females aged 15–44. In line with these aims and with identifiability considerations, our base model assumes durable immunity (no waning) with sensitivity analyses used to probe robustness of conclusions around R_0, reporting fraction, vaccine effectiveness, and campaign timing [3, 5].

3 Data and Sources

Overview. We assemble a minimal, policy-oriented dataset for South Africa covering the analysis horizon (2025 onward) and the pre-introduction validation window (2012–2019). All transformations are reproducible from the shared scripts and produce tidy CSVs used by the model.

Demography. Single-year, sex-specific population counts (`pop_age_sex.csv`) are derived from the UN World Population Prospects 2024 (WPP 2024) by filtering to South Africa and years of interest, normalising the open-ended age group (100+) to age 100, and converting units from thousands to counts. Annual births (`births.csv`) are aggregated from WPP fertility-by-age tables; an ASFR×population cross-check confirms near-equality up to rounding. Abridged life tables from WPP are expanded to single-year hazards (`mortality_singleyear.csv`) for ages 0–100 and all simulation years [4].

Pregnancy prevalence. We construct an age-specific proxy for pregnancy prevalence from WPP fertility-by-age, assuming an average pregnancy duration of 40/52 years and clipping to $[0, 1]$. We average across policy years to obtain a stable age profile (ages 15–44), consistent with WHO practice for applied rubella modelling [5].

Surveillance (validation). WHO-reported rubella cases for South Africa (2012–2019) are parsed from the public workbook into annual totals (`cases_fit.csv`). These are used for face-validity checks of magnitude and trend under literature-anchored parameters and a mean-level rescaling of the reporting fraction used for face-validity checks of magnitude and trend under literature-anchored parameters and a mean-level rescaling of the reporting fraction. We also ran an *exploratory negative-binomial likelihood fit* (Section 5.4); results informed sensitivity but were **not** used to set headline parameters. [5].

Policy inputs. A policy coverage table (`coverage.csv`) records routine coverage per year and an optional one-time catch-up campaign (coverage level, year, and age band). This table is scenario-specific and can be regenerated for alternative options during sensitivity analysis.

Reproducibility. All inputs undergo basic guardrails (completeness over years, non-negativity, unit consistency). Intermediate and final files are written to `data/processed/` and `outputs/`, and figures to `plots/`. The model reads only from these processed artefacts, ensuring a clean separation between data ingestion and simulation.

4 Model Design and Justification

4.1 State Variables and Flows (SEIR with age/sex structure)

We use an age- and sex-structured SEIR model with annual time steps, daily ODE integration within each year, and discrete demographic/vaccination events at year-end. For each sex $s \in \{\mathrm{F}, \mathrm{M}\}$ and single-year age $a \in \{0, \ldots, 100\}$ we track $S_{a,s}(t)$, $E_{a,s}(t)$, $I_{a,s}(t)$, and $R_{a,s}(t)$.

Transmission follows homogeneous mixing at the whole-population level (a conservative approximation when age-mixing matrices are unavailable), with force of infection

$$\lambda(t) \;=\; \beta \,\frac{\sum_{a,s} I_{a,s}(t)}{\sum_{a,s}\big(S_{a,s}+E_{a,s}+I_{a,s}+R_{a,s}\big)} \;+\; \varepsilon,$$

where β is the transmission rate, $\varepsilon \geq 0$ is a small importation floor, and latent/infectious progression rates are σ and γ respectively. Within-year dynamics:

$$\dot{S}_{a,s} = -\lambda S_{a,s} - \mu_{a,s}(y)\, S_{a,s}, \quad \dot{E}_{a,s} = \lambda S_{a,s} - \sigma E_{a,s} - \mu_{a,s}(y)\, E_{a,s},$$

$$\dot{I}_{a,s} = \sigma E_{a,s} - \gamma I_{a,s} - \mu_{a,s}(y)\, I_{a,s}, \quad \dot{R}_{a,s} = \gamma I_{a,s} - \mu_{a,s}(y)\, R_{a,s},$$

where $\mu_{a,s}(y)$ is the single-year mortality hazard taken from expanded abridged life tables for calendar year y. We maintain annual infection accumulators for (i) females by age, C^{F}_a, and (ii) total infections C^{tot}, via

$$\dot{C}^{\mathrm{F}}_a = \lambda\, S_{a,\mathrm{F}}, \qquad \dot{C}^{\mathrm{tot}} = \lambda \sum_{a,s} S_{a,s}.$$

This enables direct calculation of CRS and paradox diagnostics (Section 6).

4.2 Event Processes (ageing, births, routine & catch-up vaccination)

At the end of each calendar year we apply four discrete events in order: (1) *ageing* (shift each cohort up by one year, with the terminal age retaining mass), (2) *births* (add newborns to $S_{0,\mathrm{F}}$ and $S_{0,\mathrm{M}}$ by sex ratio), (3) *routine vaccination* at age 1 with effective coverage $p^{\mathrm{rout}}_y \times \mathrm{VE}$, moving susceptibles to $R_{1,s}$, and (4) an optional one-time *catch-up campaign* in year y^* with effective coverage $p^{\mathrm{catch}}_{y^*} \times \mathrm{VE}$ over a configurable age band (default 1–14 y; optional partial coverage of age 0). Births, population, and mortality follow WPP 2024; routine/catch-up coverages come from a scenario table (Section 7). This structure mirrors WHO programme practice and prior rubella models [5, 1].

4.3 Case Observation Model and CRS Estimation

Reported cases are modelled as a fraction ρ of modelled infections:

$$\widehat{\mathrm{cases}}_y \;=\; \rho\, C^{\mathrm{tot}}_y,$$

with literature-anchored ρ used for validation, and a mean-level rescaling $\hat{\rho}$ (ratio of observed to expected totals over 2012–2019) shown for face-validity; with literature-anchored ρ used for validation, and a mean-level rescaling $\hat{\rho}$ (ratio of observed to expected totals over 2012–2019) shown for face validity. In addition, we conducted an *exploratory negative-binomial likelihood fit* to (β, ρ, k) (Section 5.4); because the MLE implied heavy over-dispersion and an elevated R_0, the **main analyses retain literature-anchored parameters**, while the NB fit is reported as a sensitivity check. .

CRS is derived from infections in females 15–44 using trimester-weighted risks. Let r_1, r_2, r_3 be CRS risks for infections in trimesters 1–3 (high to low), with weights w_1, w_2, w_3 summing to 1 (approximating the distribution of gestational ages at infection). With age-specific pregnancy prevalence π_a, annual CRS is

$$\mathrm{CRS}_y \;=\; \sum_{a=15}^{44} C^{\mathrm{F}}_{a,y}\, \pi_a \left(r_1 w_1 + r_2 w_2 + r_3 w_3\right),$$

consistent with applied guidance [5]. We also track (i) cumulative infections among females 15–44 and (ii) the mean age at infection within that group, to flag paradox risk [3].

4.4 Assumptions and Parameterisation

We assume durable immunity after infection or vaccination (no waning) in the base case, consistent with rubella immunology and to preserve identifiability; sensitivity analyses probe robustness to key parameters. Vaccine effectiveness is set high (e.g., VE ≈ 0.95) following WHO [5]. The latent and infectious periods are fixed at 14 and 7 days respectively, giving $\sigma^{-1} = 14/365$ and $\gamma^{-1} = 7/365$ years. We link β to the basic reproduction number via $\beta = R_0\,\gamma$ and examine $R_0 \in \{6, 7, 8\}$ in line with the literature [1, 2, 5]. A small importation floor ε prevents numerical fade-out.

Before policy simulation, we run a multi-decade burn-in with stationary demography (fixed-year mortality, replacement births) and zero vaccination to reach endemic equilibrium, a standard approach for childhood infections [1]. Model choices (whole-population mixing, annual event timing, and CRS mapping) trade detail for transparency and are appropriate for the policy questions at hand; where relevant, we demonstrate that qualitative conclusions are insensitive to plausible parameter ranges (Sections 6–7).

5 Model Calibration & Validation

5.1 Burn-in to Endemic Equilibrium

Before policy simulation we run a multi-decade burn-in under stationary demography, zero vaccination, and literature-anchored transmissibility to reach a stable endemic state. Specifically, we hold the mortality schedule fixed at the policy-start year, use replacement births to keep population size approximately constant, and set R_0 in the plausible rubella range (base: R_0=7.5; sensitivity in Section 6) [1, 2, 5]. This provides the initial condition for all subsequent policy runs.

5.2 Literature-Anchored Parameters and Reporting

For face-valid calibration to pre-introduction surveillance, we conduct a stationary, no-vaccination run across 2012–2019 using (R_0, ρ) anchored in the literature (central values R_0=7.5, ρ=0.005), consistent with rubella's mid–single digit transmissibility and low reporting fractions [1, 2, 5]. We compare annual reported totals to model-expected totals $\widehat{\text{cases}}_y = \rho\, C^{\text{tot}}_y$.

In addition to the likelihood fit described in Section 5.4, we show a simple mean-level rescaling for face validity:

$$\hat{\rho} = \frac{\sum_{y=2012}^{2019} \text{Observed}_y}{\sum_{y=2012}^{2019} C_y^{\text{tot}}},$$

and overlay $\hat{\rho}\, C_y^{\text{tot}}$ as a reference. We emphasise that $\hat{\rho}$ is used only for validation visualisation; policy analyses retain literature-anchored parameters and explore uncertainty in Section 6. For completeness, we attempted an NB likelihood fit to annual counts (Section 5.4). Given the fitted model's heavy over-dispersion and elevated R_0, we proceeded with literature-anchored (R_0, ρ) for policy analysis.

5.3 Validation Results

Figure 1 compares observed and expected annual counts over 2012–2019. The literature-locked curve reproduces the correct order of magnitude and inter-annual variability, while the mean-level rescaled curve aligns closely with the empirical level (as intended). Given known under-ascertainment and year-to-year surveillance variation for rubella, this level of agreement supports the face validity of the chosen parameterisation for policy exploration [5].

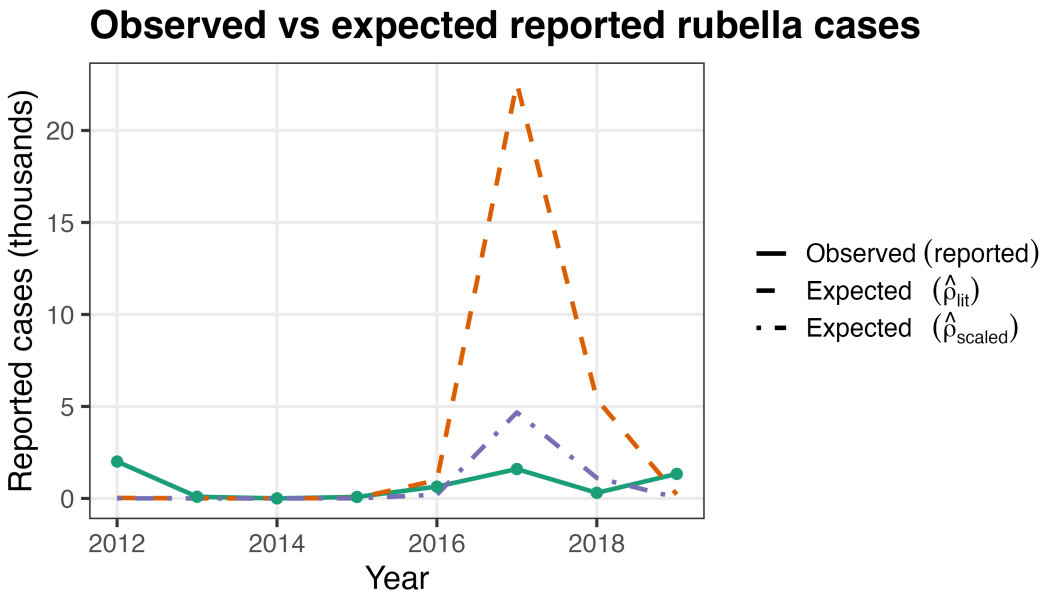

Figure 1: Observed vs expected reported rubella cases (South Africa, 2012–2019). Solid line: surveillance totals; dashed: model-expected using literature reporting fraction ρ; dot-dash: model-expected using mean-level rescaled $\hat{\rho}$.

Implications for the policy runs. The validation establishes that a literature-plausible (R_0, ρ) pair with endemic initialisation reproduces pre-introduction levels without overfitting. Hence, subsequent policy comparisons focus on changes relative to a fixed baseline (routine 80%, no catch-up) and on robustness across $R_0 \in \{6, 7, 8\}$ [1, 2, 5].

5.4 Exploratory likelihood fit (negative binomial)

We performed an exploratory fit of the transmission model to annual reported rubella cases (2012–2019) by maximizing a negative-binomial (NB) likelihood using a parallelized multi-start L-BFGS-B procedure. Let y_t be reported cases in year t and μ_t the model-predicted mean, where $\mu_t = \rho \cdot \text{infections}_t$ and ρ is the reporting fraction. We assumed

$$y_t \sim \text{NB}\big(\mu_t,\, k\big), \qquad \text{Var}(y_t) = \mu_t + \frac{\mu_t^2}{k},$$

and estimated (β, ρ, k) while holding other epidemiological parameters at literature values.

The unconstrained maximum-likelihood estimates were

$$\hat{\beta} = 577.94, \quad \hat{\rho} = 3.36 \times 10^{-3}, \quad \hat{k} = 0.395.$$

Given $\gamma = 1/(7/365) \approx 52.14\,\text{y}^{-1}$, this implies $\widehat{R_0} = \hat{\beta}/\gamma \approx 11.1$.

Why we do not adopt these as headline parameters.

- **Extreme over-dispersion.** The very small $\hat{k}$ allows the NB variance to be large enough to "forgive" substantial mean misfit; indeed, the fitted means overshoot observed points in peak years (Figure 2), indicating the likelihood is trading biological plausibility for dispersion.
- **Plausibility of R_0.** $\widehat{R_0} \approx 11$ is higher than the central rubella range typically cited for similar settings (often 5–8), and it arose here together with tiny k, a combination suggestive of weak identifiability under annual aggregation.
- **Assignment alignment.** The brief emphasizes sourcing key parameters from literature. Using literature-informed R_0 and ρ stabilizes inference and keeps the policy comparisons interpretable.

Decision. We therefore retain literature-informed R_0 (here, $R_0 = 7.5$) and ρ for the main analyses. The NB exercise is reported transparently as a sensitivity/validation check and the code is included in the Github Repo linked below.

6 Policy Scenarios & Sensitivity Analysis

6.1 Scenario Set

We evaluate routine coverage $p^{\text{rout}} \in [0.50, 0.95]$ (in 5% steps), one-time catch-up coverage $p^{\text{catch}} \in [0, 0.95]$ (in 10% steps), and timing (immediate in the policy start year, or delayed by 1–3 years). The default catch-up band is ages 1–14 y, with an alternative 9 months–9 years variant considered in scenarios. Transmission strength is varied over $R_0 \in \{6, 7, 8\}$, consistent with the literature [1, 2, 5]. The status-quo baseline for averted calculations is routine 80% with no catch-up (`r80_none`). All simulations share the endemic initial condition described in Section 5.

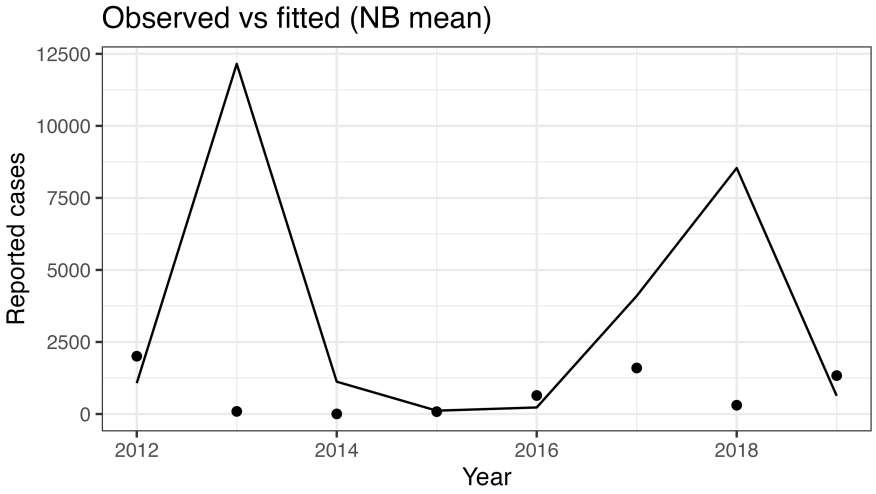

Figure 2: Observed annual reported cases (points) and fitted NB mean (line) from the unconstrained MLE. The fit achieves high likelihood via a very small k (heavy over-dispersion), while the mean trajectory shows notable peak mismatches.

6.2 Outcomes

Primary outcomes are cumulative CRS over 10 and 25 years (CRS10, CRS25). We also track (i) cumulative infections among females 15–44 (paradox check) and (ii) the mean age at infection in females 15–44 (diagnostic) [3, 5]. Averted metrics are defined as differences vs the `r80_none` baseline (e.g., $\text{Averted25} = \text{CRS25}_{\text{baseline}} - \text{CRS25}_{\text{policy}}$).

6.3 Sensitivity Grid (Routine, Catch-up, Timing)

Figure 3 shows Averted25 under an immediate campaign by routine and catch-up for each R_0. Benefits increase steeply with both routine and catch-up; at high routine ($\geq 90\%$), even moderate catch-up yields large long-horizon gains.

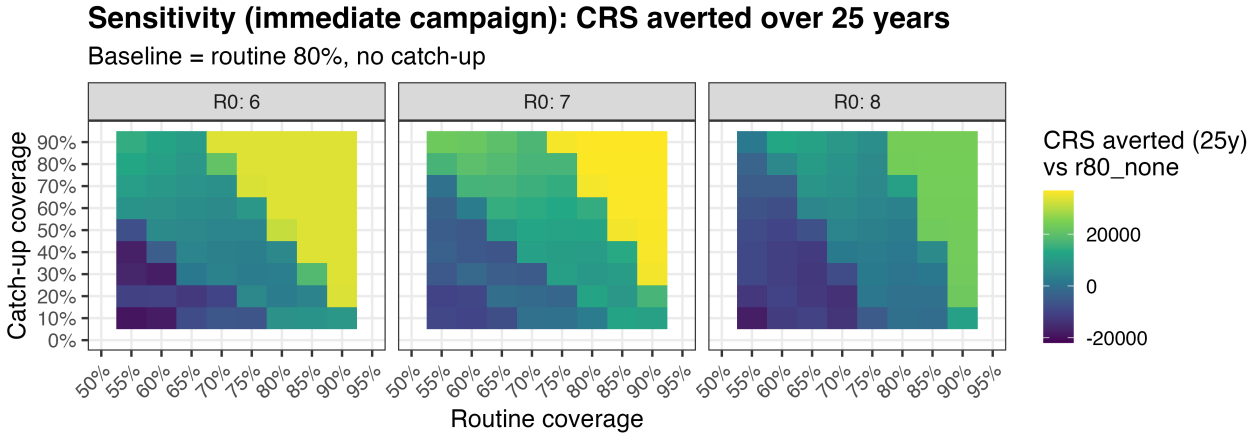

Figure 3: Sensitivity (immediate campaign): CRS averted over 25 years vs `r80_none`, by routine and catch-up; facets show $R_0 \in \{6, 7, 8\}$.

To assess safety, Figure 4 flags cells where cumulative infections in females 15–44 exceed the baseline—the vaccination "paradox" risk [3, 5]. Low routine paired with modest catch-up is most vulnerable; high routine plus broad catch-up avoids flagged regions.

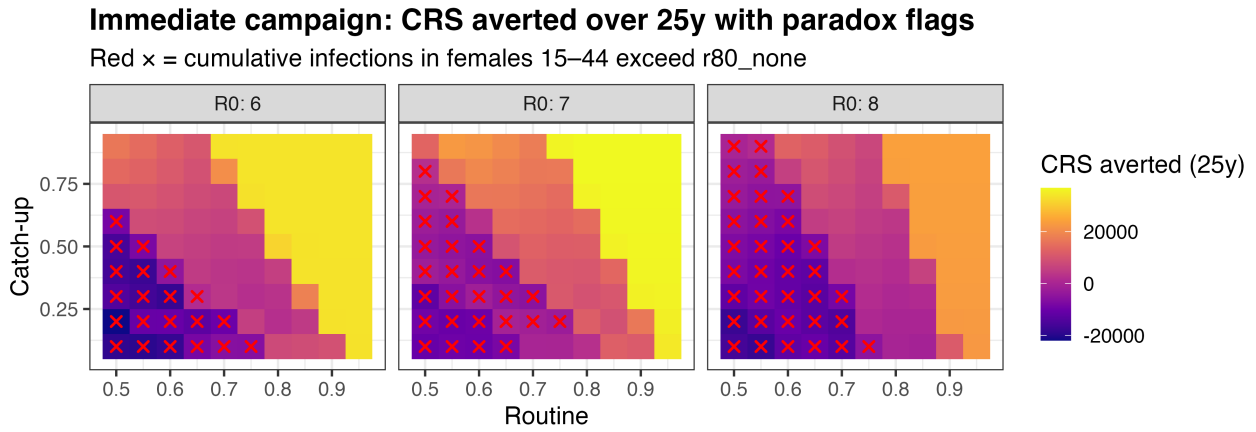

Figure 4: Immediate campaign with paradox flags: red crosses mark settings where cumulative infections among females 15–44 exceed `r80_none`.

6.4 Break-even Routine Thresholds

With a 90% catch-up (immediate), we compute the minimum routine coverage at which Averted25 > 0 vs `r80_none`. Figure 5 shows break-even near $\approx$ 55–60% for R_0=6, $\approx$ 50% for R_0=7, and $\approx$ 80% for R_0=8 (dashed lines). These thresholds summarise when introduction becomes net-beneficial over 25 years, holding other assumptions fixed.

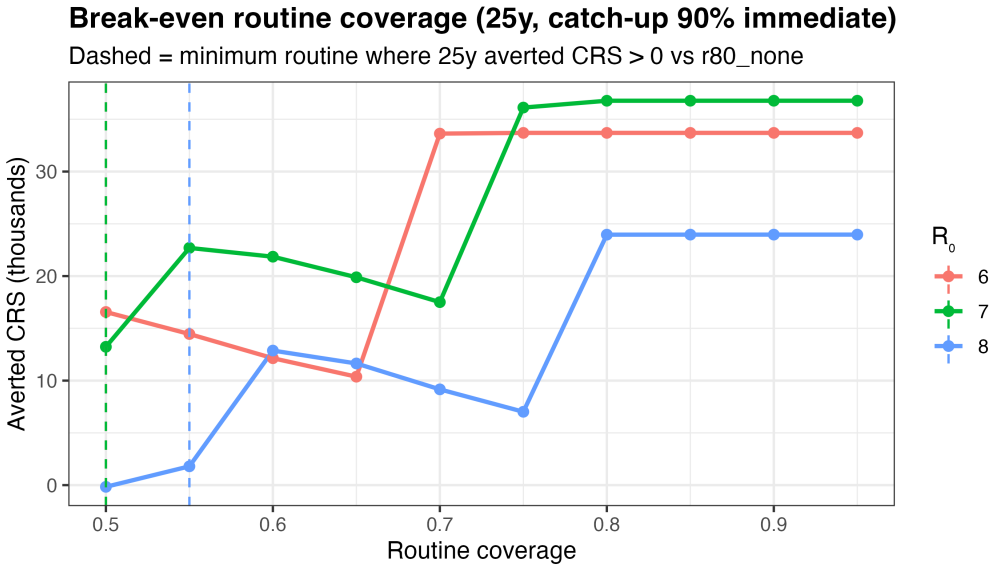

Figure 5: Break-even routine coverage (25y horizon) with 90% catch-up (immediate). Dashed lines denote the smallest routine where long-horizon averted CRS > 0.

6.5 Timing (Immediate vs Delays)

At routine 90% and catch-up 90%, Figure 6 compares immediate vs 1–3 year delays for $R_0 \in \{6, 7, 8\}$. Immediate campaigns dominate at 10 years and remain slightly better at 25 years, though long-horizon differences compress when routine is already high—a familiar result in classic rubella modelling [1, 5].

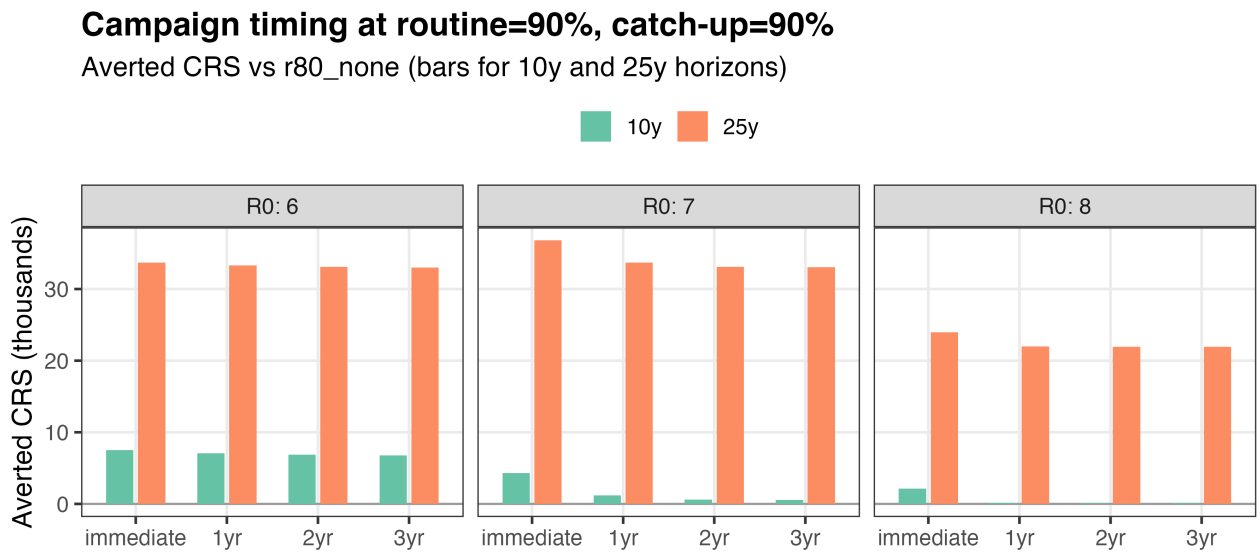

Figure 6: Campaign timing at routine 90%, catch-up 90%: averted CRS vs `r80_none` at 10y and 25y.

7 Results

7.1 Averted CRS at 10 and 25 years

Across $R_0 \in \{6, 7, 8\}$, increasing routine coverage and adding a high-coverage, wide catch-up campaign produce large reductions in CRS relative to the status-quo baseline (`r80_none`). Heatmaps for the immediate campaign (Figure 3) show a steep gradient in averted CRS as routine rises from 80% to 90–95%, with additional gains from higher catch-up coverage. At routine 90% with catch-up 90% (1–14 y), the long-horizon (25 y) benefit is substantial for all R_0 values, and the short-horizon (10 y) benefit is also large, particularly when the campaign is not delayed (Figure 6). These patterns are consistent with classic rubella dynamics where sustained reductions in force of infection are required to shift the system to a low-transmission regime [1, 5].

7.2 Break-even routine thresholds (25-year horizon)

With catch-up at 90% (immediate), the minimum routine coverage at which 25-year averted CRS becomes positive vs `r80_none` is approximately 55–60% for R_0=6, about 50% for R_0=7, and roughly 80% for R_0=8 (Figure 5). These thresholds summarise when introduction yields

a net long-run benefit, conditional on high catch-up. They also indicate that in higher-transmission settings (larger R_0), programmatic targets for routine coverage must be correspondingly higher to guarantee net benefit and to reduce paradox risk [3, 5].

7.3 Paradox check and age-at-infection diagnostic

The paradox risk—defined here as cumulative infections among females aged 15–44 exceeding the baseline—is concentrated in cells with low routine and modest catch-up (Figure 4). In these regions, vaccination reduces total infections but increases average age at infection sufficiently to raise infections in women of reproductive age, a mechanism widely discussed in the rubella literature [3, 5]. At routine $\geq$90% with catch-up 90%, the paradox flags disappear across the R_0 range examined. Consistent with this mechanism, the mean age at infection among females 15–44 rises transiently post-introduction and then declines as transmission collapses.

7.4 Timing effects (immediate vs delays)

When routine is high (90%) and catch-up is broad and well-covered (90%), implementing the campaign immediately dominates 1–3 year delays at the 10-year horizon and remains marginally better at 25 years (Figure 6). Long-horizon differences compress because sustained high routine coverage eventually closes susceptibility gaps, but immediate introduction accelerates benefits and reduces the window in which paradox mechanisms could operate [5, 3].

Summary for decision-makers. (1) High routine coverage is the principal driver of long-run CRS reduction; (2) a broad, high-coverage catch-up (e.g., 1–14 y at $\geq$90%) secures near-term gains and reduces paradox risk; (3) immediate roll-out is preferred over delays, especially at higher R_0. These conclusions are robust across the literature-plausible R_0 range [1, 2, 5].

8 Conclusion

Introducing rubella-containing vaccine (RCV) in South Africa yields large reductions in congenital rubella syndrome (CRS) when paired with high routine coverage and a broad, well-covered catch-up. Across literature-plausible transmissibility ($R_0 \in \{6, 7, 8\}$), the dominant strategy is: (i) sustain routine coverage at $\geq$ 90%; and (ii) implement a one-time catch-up across 1–14 years at $\geq$ 90% coverage. Under these conditions, near-term (10 y) and long-term (25 y) CRS averted are substantial, and the "vaccination paradox"—higher infections among females 15–44 despite fewer total infections—is avoided [3, 5].

If programme constraints imply lower routine coverage, then the minimum routine level at which 25-year benefits turn positive (vs. an 80% routine/no catch-up baseline) depends on R_0: roughly 55–60% for R_0=6, about 50% for R_0=7, and about 80% for R_0=8. In higher-transmission settings (e.g. $R_0\geq$8), introduction should not proceed without either $\geq$ 80% routine or a sufficiently broad and well-covered catch-up in the same year. Where feasible, immediate implementation (rather than delays of 1–3 years) improves 10-year outcomes and modestly improves 25-year outcomes by accelerating reductions in force of infection [1, 5].

9 Limitations & Robustness

Results reflect an age/sex-structured SEIR model with whole-population mixing, annual event timing, and trimester-weighted CRS risks. We use literature-anchored parameters with a face-valid mean-level reporting rescale for 2012–2019; we do not perform full likelihood fitting. We do not model spatial heterogeneity, seasonality, or health-system shocks, and we assume high first-dose vaccine effectiveness and durable immunity (no waning) in the base case, probing robustness across R_0, coverage levels, and timing in sensitivity analyses [1, 5]. Given under-ascertainment in rubella surveillance, our policy comparisons emphasise *relative* changes versus a fixed baseline. These simplifications are conservative for the questions at hand and align with applied rubella modelling practice [3, 5].

A Reproducibility & Code Access

All code and data prep steps are scripted. Run the files in this order from the project root:

1. `Data Preparation.Rmd`
2. `Analysis.Rmd` (*reads* processed inputs; *sources* `R/config.R` & `R/engine.R`; *writes* tables to `outputs/` and plots to `plots/`).

Full, versioned source (scripts, notebooks, and inputs) is available at: `https://github.com/Annie0619/disease_modelling_assignment`.

Repository structure

```
R/
  config.R
  engine.R
  ng_fit.R
Data Preparation.Rmd
Analysis.Rmd
data/
    processed   # All outputs from Data Preparation.Rmd
plots/          # Figures exported from Analysis.Rmd
```

Reproduce in two commands

```
Rscript -e "rmarkdown::render('Data Preparation.Rmd')"
Rscript -e "rmarkdown::render('Analysis.Rmd')"
```

B Data preprocessing

This appendix documents the preparation of demographic, fertility, vaccination, surveillance, pregnancy–prevalence, and mortality inputs for the South Africa rubella and congenital rubella syndrome (CRS) analysis. Demographic, fertility, and mortality inputs were obtained from the United Nations World Population Prospects 2024 (WPP) standard projections (CSV format) [4]. Surveillance inputs were obtained from the World Health Organization (WHO) Immunization Data portal (provisional measles and rubella) [6]. All processing aimed to produce internally consistent inputs aligned to a prospective ten–year policy window, with *single–year ages* used throughout.

B.1 Population by age and sex

Source. WPP 2024 population by single–year age and sex for South Africa [4].
Procedure. Records were filtered to South Africa and the analysis years. Ages recorded as open–ended ("100+") were standardised to 100. WPP populations reported in thousands were converted to counts. The resulting tidy structure comprises year, single–year age (0–100), sex (F/M), and population.

Rationale. The age–sex structure is required to initialise SEIR compartments, apply annual ageing (one year per calendar step), and provide denominators for incidence and CRS.
Assumption. The Constant–fertility projection variant was used for coherence with fertility inputs.

B.2 Annual live births

Source. WPP 2024 fertility by single–year age of mother (including births by maternal age and ASFR) [4].
Procedure. Data were filtered to South Africa, the analysis years, and the same projection variant as the population. Annual live births were obtained by summing births across maternal ages (15–49, with 50 retained where present but typically negligible) and converting from thousands to counts. For diagnostic purposes, births were cross–checked against ASFR multiplied by female population at each age; close agreement confirmed internal consistency.
Rationale. Annual births provide the inflow to the age–0 class at year boundaries. The inflow is split by sex using a conventional sex ratio at birth when applied in the model.
Assumption. Alignment of population and fertility variants to avoid internal inconsistencies.

B.3 Vaccination coverage (scenario input)

Source. Scenario specification by year; WUENIC may be used for contextual displays but is not required for the model run.
Procedure. A year–level table was constructed for the policy window specifying routine coverage (applied at approximately age 1 each year) and a one–time catch–up coverage in the designated campaign year for ages 1–14. Values are proportions in $[0, 1]$ and can be varied across scenarios.
Rationale. Explicit coverage schedules are required to implement the routine vaccination pulse at age 1 and the one–time catch–up pulse at year boundaries.
Assumptions. Equal coverage by sex; vaccine effectiveness applied as direct movement from susceptible to recovered at the pulse; no waning in the base case.

B.4 Reported rubella cases

Source. WHO Immunization Data portal, provisional measles and rubella dataset [6].
Procedure. The annual rubella table was harmonised for variable names and filtered to South Africa. Where multiple confirmation categories were present, lab–confirmed rubella counts were used; if lab–confirmed equalled total confirmed, either measure was acceptable provided the choice was consistent across years. A primary fitting window of 2012–2019 was adopted to capture a stable pre–COVID, pre–introduction period; partial or atypical years were excluded from the primary fit.
Rationale. A consistent annual time series of reported cases is used for face-validity checks of magnitude and trend under literature-anchored parameters; we also show a mean-level rescaling of the reporting fraction. No likelihood fitting was performed.
Assumption. Surveillance definitions are sufficiently stable within the fitting window; deviations are addressed by window selection and sensitivity analysis.

B.5 Pregnancy prevalence by age

Source. Derived from WPP 2024 ASFR and female population by age [4].
Procedure. For each female age a, point prevalence of pregnancy was approximated as $\text{ASFR}(a) \times D$, with $D = 40/52 \approx 0.77$ years (average pregnancy duration). When ASFR was reported per 1,000 women, a conversion to per–woman units preceded the calculation. The resulting age–specific prevalence was averaged over the analysis years to obtain a stable single–year profile for ages 15–44 and bounded to $[0, 1]$.
Rationale. Age–specific pregnancy prevalence is required to convert female infections (ages 15–44) into expected CRS cases using trimester–specific risk parameters.
Assumptions. Uniform distribution of conceptions across the year and average gestation length; early losses are not explicitly upweighted in the base case and are explored in sensitivity analysis.

B.6 All–cause mortality by single–year age and sex

Source. WPP 2024 life tables (abridged) for South Africa, Medium variant [4].
Procedure. The abridged life table provides, for each age band $[a, a + n)$, the one–year death probability q_x and the central death rate m_x. A continuous–time annual hazard was constructed for each band as

$$\mu_{\text{band}} \;=\; -\ln\bigl(1 - q_{\text{band}}\bigr) \,/\, n,$$

where n is the age–band span in years. This band–level hazard was then expanded to *single–year ages* by assigning the same μ to each age within the band; the open–ended 100+ group was mapped to age 100. The result is a tidy table of hazards by year, single–year age (0–100), and sex. In the baseline specification, hazards are held fixed at the start year across the ten–year horizon; a time–varying specification is straightforward if desired.
Rationale. Continuous hazards allow deaths to be modelled within the ODE alongside infection dynamics, ensuring coherent cohort accounting over the year.
Assumptions. Piecewise–constant hazards within abridged age bands; mapping of 100+ to age 100; baseline hazards held constant over the policy window in the primary analysis.

B.7 General assumptions and quality checks

Single–year ages (0–100) are used universally for clarity and exact age targeting (routine at age 1, catch–up 1–14, CRS 15–44, and mortality). Unit conversions from thousands to counts were applied wherever necessary. Open–ended ages were standardised for reproducibility. Tidy, standardised fields (year, age, sex, and measure) were enforced across inputs. Consistency checks compared births derived from reported births versus ASFR×population and verified that single–year totals match aggregated diagnostics.

B.8 Use of sources

WPP 2024 provided population structure, fertility schedules, and life tables because it offers internationally harmonised, age–specific estimates and projections suitable for mechanistic

modelling [4]. The WHO provisional rubella dataset provided country–level, annually collated case totals sufficient for fitting an observation model [6]. Citations to these sources appear in the main text and this appendix.

References

[1] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, UK, 1991. Classic text; rubella R_0 typically reported in the mid–single digits.

[2] W. John Edmunds, Nigel J. Gay, Mirjam Kretzschmar, Richard G. Pebody, and Heather Wachmann. The pre-vaccination epidemiology of measles, mumps and rubella in england and wales: implications for modelling studies. *Epidemiology and Infection*, 125(3):635–650, 2000. Includes empirical estimates of rubella R_0 in a high-surveillance setting.

[3] C. Jessica E. Metcalf, Justin Lessler, Petra Klepac, Felicity T. Cutts, and Bryan T. Grenfell. Implications of measles elimination for rubella control and congenital rubella syndrome. *The Lancet*, 379(9811):1187–1188, 2012. Short communication discussing rubella control, elimination targets, and the vaccination paradox.

[4] United Nations, Department of Economic and Social Affairs, Population Division. World population prospects 2024: Standard projections (csv format). `https://population.un.org/wpp/downloads?folder=Standard%20Projections&group=CSV%20format`, 2024. Accessed 2025-10-16.

[5] World Health Organization. Rubella vaccines: Who position paper – july 2020. *Weekly Epidemiological Record*, 95(27):306–324, 2020. Summarises epidemiology, vaccination strategy, and typical transmissibility parameters.

[6] World Health Organization. Provisional measles and rubella data. `https://immunizationdata.who.int/global?topic=Provisional-measles-and-rubella-data&location=`, 2025. Accessed 2025-10-16.