serocalculator, an R package for estimating seroincidence from cross-sectional serological data

Authors: Kristina W. Lai[1]*, Chris Orwa[2], Jessica C. Seidman[3], Denise O. Garrett[3], Samir K. Saha[4], Dipesh Tamrakar[5], Farah Naz Qamar[6], Richelle C. Charles[7,8], Jason R. Andrews[9], Peter Teunis[10], Kristen Aiemjoy[1,11]†, Douglas Ezra Morrison[1]†

Affiliations:

1. Department of Public Health Sciences, University of California Davis School of Medicine, Davis, CA, USA
2. SkyeHi Technologies, Nairobi, Kenya
3. Sabin Vaccine Institute, Washington, DC, USA
4. Child Health Research Foundation, Dhaka, Bangladesh
5. Dhulikhel Hospital, Kathmandu University Hospital, Dhulikhel, Nepal
6. Department of Paediatrics and Child Health, Aga Khan University, Karachi, Pakistan
7. Harvard Medical School, Harvard University, Boston, MA, USA
8. Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Harvard University, Boston, MA, USA
9. Division of Infectious Diseases and Geographic Medicine, Stanford University School of Medicine, Stanford, CA, USA
10. Hubert Department of Global Health, Center for Global Safe WASH, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
11. Department of Immunology and Microbiology, Mahidol University Faculty of Tropical Medicine, Bangkok, Thailand

***Corresponding author:**

Kristina W. Lai
Medical Sciences 1C
One Shields Avenue
Davis, CA 95616
kwlai@ucdavis.edu

†Joint senior authors

Word Count: 1987 words

Abstract: (150/150 words)

Motivation: Seroincidence—the rate of new infections in a population—is a key measure for understanding pathogen transmission dynamics and informing public health action. Estimating seroincidence from cross-sectional data is complicated by antibody waning, cross-reactivity, and individual heterogeneity in antibody responses.

Implementation: serocalculator is an open-source R package that uses a likelihood-based framework incorporating antibody decay models, biological variability, and measurement noise to estimate seroincidence rates from cross-sectional serological data.

General features: The package supports overall and stratified seroincidence estimation using single or multiple biomarkers. It requires three inputs: (1) a pre-estimated seroresponse model characterizing post-infection antibody waning; (2) noise parameters capturing biological and assay-related variability; and (3) quantitative antibody responses from a cross-sectional survey. It is computationally efficient, well-documented, and includes a point-and-click R Shiny interface. These features promote usability across research and public health.

Availability: The package **serocalculator** is freely available on CRAN, with development versions on GitHub.

Introduction

Serological surveys can be used to measure population-level antibody responses to characterize infection occurrence (1,2). Following pathogen exposure, the adaptive immune system generates antigen-specific antibodies, which increase rapidly after infection and gradually wane over time (1,3,4). The concentration of antibodies against a given pathogen reflects the time since last infection (2–4). In cross-sectional surveys, a large proportion of individuals with high antibody concentrations suggests high burden of recent or frequent infection, while a smaller proportion indicates lower burden.

Population-level antibody responses are typically characterized using two key epidemiological measures: seroprevalence, the proportion of individuals with antibody levels above a defined threshold; and seroincidence, the rate at which new infections occur in the population (1,2). The magnitude and speed of an individual's immune response to infection is complex and dynamic, shaped by infectious dose, age, disease severity, prior exposures, and vaccination history. Most analytic methods and software used to estimate seroprevalence and seroincidence dichotomize quantitative antibody levels into binary outcomes—seropositive or seronegative. While simple and widely used, this approach ignores temporal changes in antibody concentrations, individual-level heterogeneity, and immunological cross-reactivity between pathogens. Moreover, seroprevalence reflects the cumulative history of prior exposures in a population and is influenced by the age distribution of the sample, making it an imperfect proxy for recent transmission. In contrast, seroincidence rates quantify the rate at which new infections occur and are increasingly favored by epidemiologists, modelers, and public health professionals (3,4). These rates can help determine whether a disease has been eliminated (5,6), identify risk factors for transmission, and locate areas where interventions are most needed (3,7).

Traditional methods to assess seroincidence require resource- and time-intensive longitudinal studies that are often infeasible. However, new statistical methods estimate seroincidence rates from cross-sectional serosurveys by incorporating models of antibody decay dynamics from confirmed cases (2,8,9). Seroincidence can be estimated as a function of peak antibody response and decay rate from infected individuals while accounting for heterogeneity in immunity, responses to multiple antigens, age-dependence, non-specific binding of the assay probes (noise), and laboratory assay variability (1,8–10). This approach assumes that exposure follows a Poisson process, meaning all individuals in the sampled population are subject to a constant infection risk over time. Using this assumption, maximum likelihood estimation (MLE) is used to estimate seroincidence from cross-sectional serologic data (2,8–11). We also assume that seroresponses from each antibody isotype are independent, allowing us to sum the likelihoods from multiple antibodies and antigens, thereby increasing precision in the seroincidence estimate (1,9). Here, we introduce **serocalculator**, a statistical software package that offers a validated and accessible framework for estimating seroincidence rates from cross-sectional serosurveys by integrating antibody kinetics with noise correction to produce dynamic infection rate estimates (1,8–10). It includes intuitive tools such as multi-biomarker support, stratified analysis, and an R Shiny interface for non-coders. An earlier package, *seroincidence* (12), was developed and maintained from 2015-2018, but lacked methods for incorporating biologic and measurement noise, necessitating a full revision. The new **serocalculator** package supersedes *seroincidence* by incorporating the latest methodological advancements for seroincidence estimation.

Implementation

serocalculator can be downloaded from the Comprehensive R Archive Network (CRAN). Over 30 exported functions guide users from data import and cleaning to analysis and visualization.

All functions in **serocalculator** version 1.4.0 are described on the package website. The general workflow is: 1) import data, 2) inspect and visualize data, and 3) estimate seroincidence rate.

Import Data

Three inputs are required: 1) Previously-estimated longitudinal seroresponse model parameters characterizing the longitudinal response of the surveyed serum antibodies; 2) Noise parameters that encompass biological noise (individual variation in response) and measurement noise (laboratory assay variability); and 3) Quantitative antibody levels from a cross-sectional population-based serosurvey sample that is calibrated on the same scale as the longitudinal seroresponse parameters. For testing purposes, users can access small example files within the package, or they can import larger sample files for specific diseases from the publicly available Serocalculator Data Repository on Open Science Framework. Example input datasets and formatting guidance are provided in Supplement S1.

Longitudinal Seroresponse Parameters

Longitudinal seroresponse parameters describe how antibodies rise and fall after infection. Before using **serocalculator** to estimate seroincidence, these parameters must be previously estimated by fitting two-phase, hierarchical within-host models to quantitative antibody responses measured longitudinally in a cohort of confirmed cases, using a Bayesian inferential framework and a Markov Chain Monte Carlo (MCMC) computational approach to sample from the joint predictive posterior distribution of the model parameters for each antigen isotype (1,8–10) (see Supplement S2). Seroresponse models can be stratified by age, country, and/or other relevant demographic characteristics, if seroresponse dynamics vary among the corresponding subpopulations.

Users can import their own set of seroresponse parameters in either CSV or RDS format and then apply `as_sr_params()`, which will assign the imported dataset as an `sr_params` object and ensure required variables are present and in the correct format. If using online files from a URL, users can apply the function `load_sr_params()`, which will perform necessary column and formatting checks. The longitudinal parameter sets must include the following columns: one or more disease-specific antigen and antibody isotype pairs that matches those included in the cross-sectional population data (`antigen_iso`), the baseline antibody concentration (`y0`), the peak antibody concentration (`y1`), the time to peak antibody concentration (`t1`), the antibody decay rate (`alpha`), and the antibody decay shape (`r`).

Further details on the longitudinal seroresponse parameters can be found in Supplement S2.

Noise Parameters

The noise parameters object defines assumptions about the measurement process.(1) Biological noise (ν) reflects potential non-specific antibody binding from other exposures, such as cross-reactivity with a related pathogen. It is defined as the 95th percentile of the antibody response distribution, assumed to be normal, in a population with no prior exposure to the pathogen.(1) Measurement noise (ϵ) represents laboratory assay variability from the laboratory testing process. It is defined by a coefficient of variation (CV), or the ratio of the standard deviation to the mean for replicates across plates. The **serocalculator** package can also accommodate censored data: lower and upper limits of reliable quantification may be specified as y_{low} and y_{high}, respectively, and any observations recorded outside of those limits will be treated as censored.

As above, users can import their own noise parameters into R (e.g., from CSV or RDS files) as a `data.frame` and then apply `as_noise_params()`, which will convert the `data.frame` into a

`noise_params` object (a subclass of `tibble::tbl_df` with added metadata attributes) and ensure required variables are present and in the correct format. If using online files from a URL, users can apply the function `load_noise_params()` and then follow the process described above for the seroresponse parameters. The required noise parameter variables are: one or more disease-specific antigen and antibody isotype that match those included in the cross-sectional population data (`antigen_iso`), the biological noise (`nu`), the measurement noise (`eps`), and the lower and upper limits of detection of the assay (`y.low` and `y.high` respectively). Further details on the noise parameter calculations are in Supplement S3.

Cross-sectional population data

Once again, users can import their own cross-sectional population data in either CSV or RDS format and apply the function `as_pop_data()`, which assigns the imported dataset as a `pop_data` object and ensures correct formatting. The required variables for this dataset are: age, measured in years (`age`), one or more specified antigen isotype pairs (`antigen_iso`), quantitative antibody response (`value`), and a unique participant id (`id`). The `as_pop_data()` function detects the required variables or allows users to specify them. If using online data from a URL for an online repository such as Open Science Framework, users can apply the function `load_pop_data()` and then once again follow the process described above for the seroresponse parameters. Population datasets can include additional variables for stratification.

Estimating seroincidence

Once all parameters and datasets are imported, users can estimate either overall or stratified seroincidence rates using the functions `est_seroincidence()` and `est_seroincidence_by()`, respectively. Users begin by specifying the three required data inputs and selecting one or more antigen-isotype pairs for inclusion in the estimate. Optional parameters allow customization of

visualizations, warning messages, parallel computing, and the estimation process. Stratified seroincidence estimates can be generated by specifying one or more covariates in the cross-sectional dataset, and optionally in the longitudinal seroresponse parameters.

Use

We will demonstrate the use of **serocalculator** with a reproducible example of enteric fever seroincidence using a subset of data from a series of serosurveys conducted in Bangladesh, Nepal and Pakistan in 2016-2021 (13,14). Additional details on these data are available in Supplement S4. Data for this analysis are stored in the Serocalculator Data Repository on Open Science Framework. This example is also available as an article on the **serocalculator** website.

First, we load each of our three inputs (cross-sectional population data, noise parameters, and seroresponse parameters) and then create descriptive plots for both cross-sectional data and modeled seroresponse parameters using built-in `autoplot()` visualization methods. In this example, we import noise parameters calculated at each study site laboratory from analysis using enzyme-linked immunosorbent assays (ELISA).

[Figure 1]

We can estimate overall and stratified seroincidence rates using the `est_seroincidence()` and `est_seroincidence_by()` functions, respectively. In both cases, we combine IgA and IgG seroresponses to the hemolysin E (HlyE) antigen (Figure 2a). For the overall seroincidence rate, we choose to filter to only sites in Pakistan and find a seroincidence rate of 128 new infections per 1000 person-years (95% CI). To compare across countries and ages, we specify the strata variables and find that 5–15-year-olds in Bangladesh experience the highest seroincidence of

enteric fever, with a rate of 477 new infections per 1000 person-years (Figure 2). Model outputs from `est_seroincidence()` and `est_seroincidence_by()` can be summarized and visualized using the corresponding `summary()` and `autoplot()` methods, respectively.

[Figure 2]

Discussion

serocalculator provides a rapid and computationally simple method for calculating population-level seroincidence rates that account for antibody decay dynamics, biologic variability, and measurement noise and can integrate information from multiple biomarkers. Building upon the prior *seroincidence* package (12), the ***serocalculator*** package implements updated analytic methods that account for biological and measurement noise, new function names, new summary and visualization options, and cross-sectional data simulations.

Several groups have developed R packages to support seroepidemiologic analysis. The *serosv* package estimates seroprevalence and force of infection (FOI) using built-in datasets, but it is not intended for user-supplied data (15). *serosim* simulates serological study data to support study design and evaluate within-host dynamics and new analytic approaches (16). While **serocalculator** also includes a simulation function, its primary purpose is to estimate seroincidence rates from user-collected data. *Rsero* provides Bayesian tools for comparing serocatalytic models, assessing FOI under various assumptions (e.g., transmission mode, seroreversion), and evaluating sampling strategies (17). *serojump* focuses on infection timing and antibody kinetics using Bayesian reversible jump MCMC and individual-level data (18). Collectively, these packages expand access to seroepidemiologic tools. Notably, **serocalculator** is the only package to incorporate individual-level heterogeneity, combine multiple antigen isotypes, and account for multiple sources of noise.

Future enhancements will address current methodological limitations and expand usability. While **serocalculator** supports interval censoring for some assays, its application is constrained by the availability of longitudinal seroresponse parameters for specific pathogens and assays. To increase flexibility across study settings, we are developing a complementary R package that will enable users to model seroresponse parameters from their own data.

Another key development area is improving performance in high-burden settings, where repeated exposures can bias seroincidence estimates by changing the distribution of longitudinal seroresponse patterns. Current methods do not fully address this challenge, and we are actively working to incorporate Approximate Bayesian Computation techniques to enhance estimation in areas with high force of infection (11).

The **serocalculator** package has been applied to studies of enteric fever and scrub typhus (14,19,20), and its methods are currently being adopted by research teams and public health institutions to inform vaccine introduction decisions and surveillance strategies in endemic regions. **serocalculator** provides a free, open-source, and user-friendly tool for estimating seroincidence rates from cross-sectional quantitative serology data. Additional documentation and tutorials are available at https://ucd-serg.github.io/serocalculator/.

Funding:

This work was supported by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health [grant number R21AI176416 to KA and DEM; grant number F31AI194664 to KWL] and the Fogarty International Center at the National Institutes of Health [grant number K01TW012177 to KA].

Conflict of interest:

None declared.

Supplementary data:

Supplementary material are available online.

Data availability:

Source code for the **serocalculator** package is available on GitHub: https://github.com/UCD-SERG/serocalculator. Deidentified data from the sample analysis is hosted on Open Science Framework: https://osf.io/ne8pc/files/osfstorage.

References

1. Teunis PFM, van Eijkeren JCH. Estimation of seroconversion rates for infectious diseases: Effects of age and noise. Stat Med [Internet]. 2020 Sep 20;39(21):2799–814. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8578
2. Simonsen J, Mølbak K, Falkenhorst G, Krogfelt KA, Linneberg A, Teunis PFM. Estimation of incidences of infectious diseases based on antibody measurements. Stat Med [Internet]. 2009 Jun 30;28(14):1882–95. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.3592
3. Aiemjoy K, Seidman JC, Charles RC, Andrews JR. Seroepidemiology for Enteric Fever: Emerging Approaches and Opportunities. Open Forum Infect Dis. 2023 Jun 2;10(Supplement_1):S21–5.
4. Hay JA, Routledge I, Takahashi S. Serodynamics: A primer and synthetic review of methods for epidemiological inference using serological data. Epidemics. 2024 Dec;49:100806.
5. Kamau E, Ante-Testard PA, Gwyn S, Blumberg S, Abdalla Z, Aiemjoy K, et al. Characterizing trachoma elimination using serology. 2024.
6. Tedijanto C, Solomon AW, Martin DL, Nash SD, Keenan JD, Lietman TM, et al. Monitoring transmission intensity of trachoma with serology. Nat Commun. 2023 Jun 5;14(1):3269.
7. Cutts FT, Hanson M. Seroepidemiology: an underused tool for designing and monitoring vaccination programmes in low- and middle-income countries. Tropical Medicine & International Health. 2016 Sep;21(9):1086–98.

8. de Graaf WF, Kretzschmar MEE, Teunis PFM, Diekmann O. A two-phase within-host model for immune response and its application to serological profiles of pertussis. Epidemics. 2014 Dec;9:1–7.

9. Teunis PFM, van Eijkeren JCH, Ang CW, van Duynhoven YTHP, Simonsen JB, Strid MA, et al. Biomarker dynamics: Estimating infection rates from serological data. Stat Med. 2012 Sep 10;31(20):2240–8.

10. Teunis PFM, van Eijkeren JCH, de Graaf WF, Marinović AB, Kretzschmar MEE. Linking the seroresponse to infection to within-host heterogeneity in antibody production. Epidemics [Internet]. 2016 Sep;16:33–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1755436516300135

11. Teunis PFM, Wang Y, Aiemjoy K, Kretzschmar M, Aerts M. Estimating seroconversion rates accounting for repeated infections by approximate Bayesian computation. Stat Med. 2023 Dec 10;42(28):5160–88.

12. Teunis P, van Eijkeren J. Estimating Infection Rates from Serological Data. R package version 2.0.0. Archived May 23, 2022 [Internet]. 2018. Available from: https://CRAN.R-project.org/package=seroincidence

13. Garrett DO, Longley AT, Aiemjoy K, Yousafzai MT, Hemlock C, Yu AT, et al. Incidence of typhoid and paratyphoid fever in Bangladesh, Nepal, and Pakistan: results of the Surveillance for Enteric Fever in Asia Project. Lancet Glob Health. 2022 Jul;10(7):e978–88.

14. Aiemjoy K, Seidman JC, Saha S, Munira SJ, Islam Sajib MS, Sium SM Al, et al. Estimating typhoid incidence from community-based serosurveys: a multicohort study. Lancet Microbe. 2022 Aug 1;3(8):e578–87.

15. Phan Truong Quynh A, Pham Nguyen The N, Bui Thanh L, Huynh T, Ong T, Choisy M. serosv: Model Infectious Disease Parameters from Serosurveys. [Internet]. 2025 [cited 2025 Jan 29]. Available from: https://oucru-modelling.github.io/serosv/

16. Menezes A, Takahashi S, Routledge I, Metcalf CJE, Graham AL, Hay JA. serosim: An R package for simulating serological data arising from vaccination, epidemiological and antibody kinetics processes. PLoS Comput Biol. 2023 Aug 14;19(8):e1011384.

17. Hozé N, Pons-Salort M, Metcalf CJE, White M, Salje H, Cauchemez S. RSero: A user-friendly R package to reconstruct pathogen circulation history from seroprevalence studies. PLoS Comput Biol. 2025 Feb 3;21(2):e1012777.

18. Hodgson D, Hay J, Jarju S, Jobe D, Wenlock R, de Silva TI, et al. serojump: A Bayesian tool for inferring infection timing and antibody kinetics from longitudinal serological data. 2025.

19. Aiemjoy K, Katuwal N, Vaidya K, Shrestha S, Thapa M, Teunis P, et al. Estimating the Seroincidence of Scrub Typhus using Antibody Dynamics after Infection. Am J Trop Med Hyg. 2024 Aug 7;111(2):267–76.

20. Aiemjoy K, Rumunu J, Hassen JJ, Wiens KE, Garrett D, Kamenskaya P, et al. Seroincidence of Enteric Fever, Juba, South Sudan. Emerg Infect Dis. 2022 Nov;28(11).

Figure captions

Figure 1: Code and visualization of log antibody responses from three study sites. a) Code for importing the three required inputs. b) Code for visualizing cross-sectional antibody data and longitudinal seroresponse. c) Plot for cross-sectional curve parameters. d) Plot for longitudinal seroresponse parameters.

Alt Text: R code and output depicting data import and visual inspection in subfigures labeled from a to d.

Figure 2: Estimation of seroincidence rates stratified by country and age category. a) Code for executing and summarizing results of the `est_seroincidence_by()` function. b) Output from `summary()` for stratified seroincidence estimates. c) Code for plotting seroincidence rate estimates by country and age category as a bar plot. d) Output from `autoplot()` for visualizing seroincidence by strata.

Alt Text: R code and results showing seroincidence estimation, results summary, and visualization in subfigures labeled from a to d.

R code

a) Import data

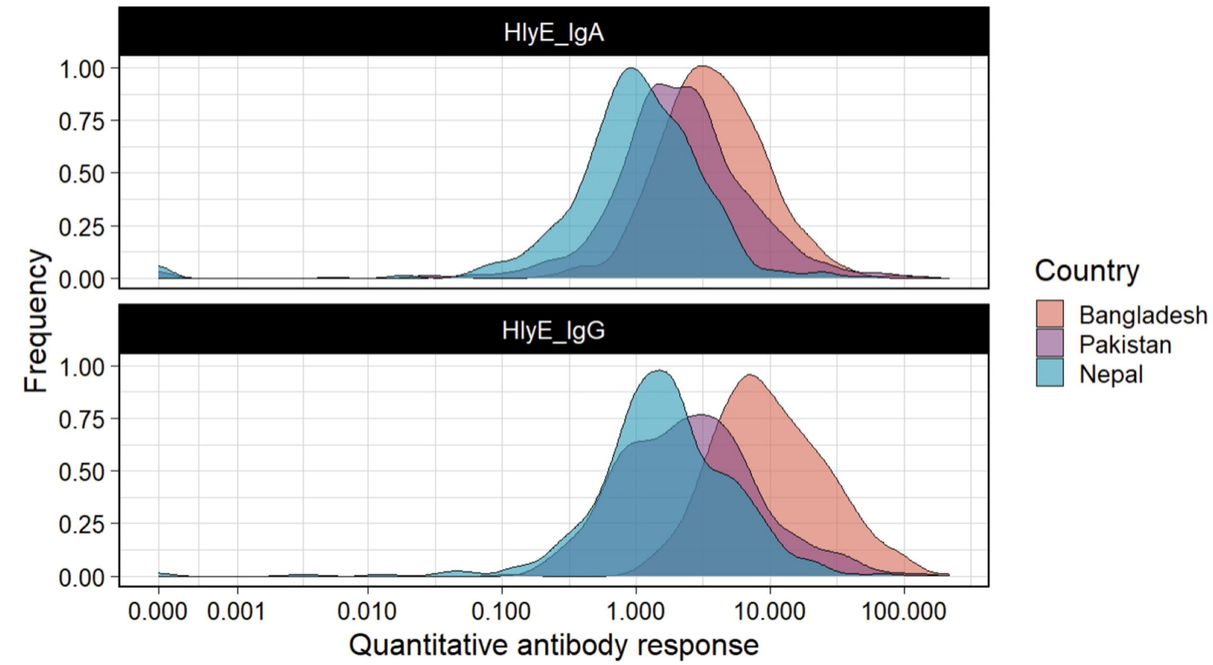
```
# Import sample cross-sectional data
pop <- "https://osf.io/download//n6cp3/" |>
  load_pop_data()

# Import longitudinal seroresponse parameters
sr <- "https://osf.io/download/rtw5k/" |>
  load_sr_params()

# Import sample noise parameters
noise <- url("https://osf.io/download//hqy4v/") |>
  load_noise_params()
```

b) Visualize data

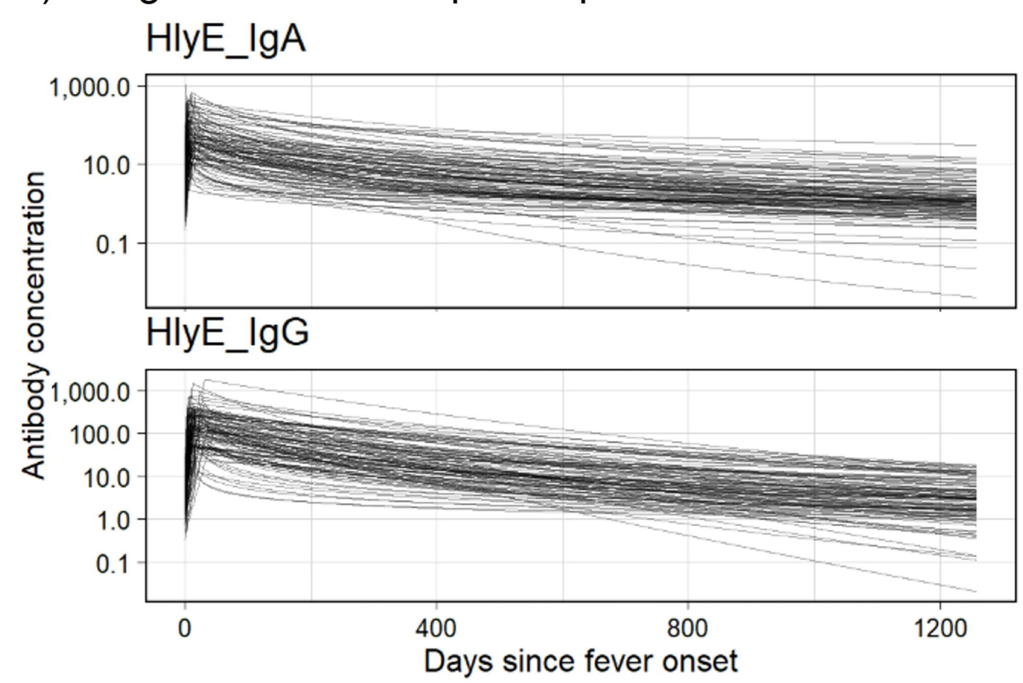
```
# Visualize cross-sectional antibody data
xs_data |>
  autoplot(strata = "Country", type = "density",
log = TRUE)

# Visualize seroresponse parameters
curves |>  filter(antigen_iso %in% c("HlyE_IgA",
"HlyE_IgG")) |>
  autoplot()
```

Output

c) Cross-sectional antibody data



d) Longitudinal seroresponse parameters



R code

a) Estimate stratified seroconversion rate

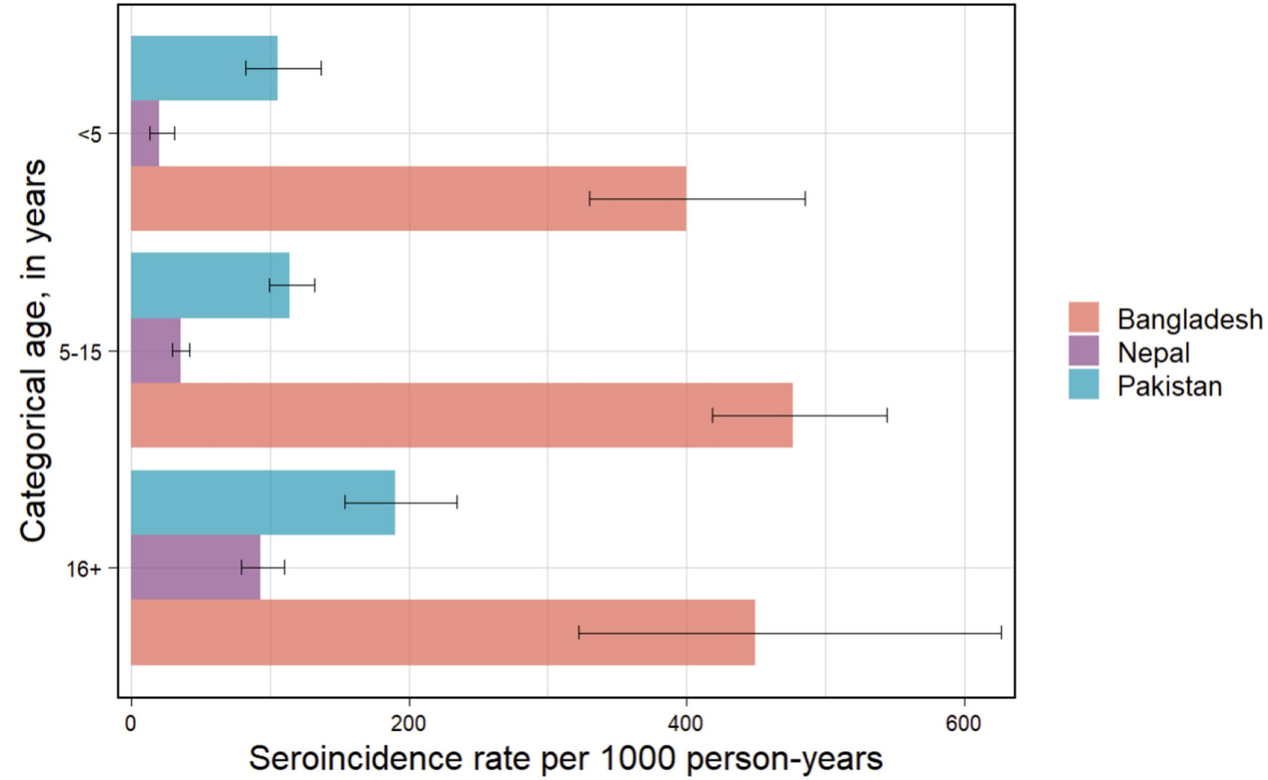
```
# Estimate seroconversion rate
est_country_age = estimate_scr_by(
   strata = c("Country", "ageCat"),
   pop_data = pop,
   sr_params = sr,
   noise_params = noise,
   antigen_isos = c("HlyE_IgG", "HlyE_IgA")
)

# Summarize results and save as dataframe
est_country_agedf <-
summary(est_country_age)
```

c) Visualize stratified seroconversion rate

```
# Plot seroconversion rates as a bar plot
autoplot(
   est_country_agedf,
   type = "bar",
   yvar = "ageCat",
   color_var = "Country",
   CIs = TRUE
)
```

Output

b) Abbreviated summary output

Stratum	Country	ageCat	n	incidence.rate	SE	CI.lwr	CI.upr
Stratum 1	Bangladesh	<5	101	0.39998293	0.0395	0.3297	0.4853
Stratum 2	Bangladesh	5-15	256	0.47701125	0.032	0.4183	0.544
Stratum 3	Bangladesh	16+	44	0.44929893	0.0763	0.3221	0.6267
Stratum 4	Nepal	<5	171	0.02026628	0.0044	0.0132	0.0311
Stratum 5	Nepal	5-15	378	0.0354936	0.0031	0.0299	0.0421
Stratum 6	Nepal	16+	211	0.0935101	0.0078	0.0795	0.11
Stratum 7	Pakistan	<5	126	0.10592089	0.0136	0.0823	0.1363
Stratum 8	Pakistan	5-15	261	0.1145304	0.0084	0.0991	0.1323
Stratum 9	Pakistan	16+	107	0.19011951	0.0204	0.1541	0.2346

d) Visualization