

## 5. Survey analysis and reporting

### 5.1. Data management

Serosurvey projects require substantial data management with numerous components including:

- Collecting, storing, entering and cleaning datasets obtained from survey questionnaires;
- Tracking the location, cataloguing and storing specimens;
- Linking and merging laboratory test results with the survey dataset;
- Identifying which laboratory results to include in seroprevalence calculations;
- Process monitoring of the quality of fieldwork (e.g. GPS data).

See **Section 4.5 for interpretation of test results.**

#### 5.1.1. Variables to describe complex sample and survey weights

It is important to enlist the assistance of a sampling statistician to oversee the sample design and sample selection, track the information required to calculate survey weighting, and conduct mixture modelling of laboratory data if appropriate.

To make appropriate population level estimates of seroprevalence and to estimate meaningful prevalence confidence intervals, it is necessary to use estimation methods that incorporate survey weights and account for the complex nature of the survey sample. Several modern software packages handle these calculations including EpiInfo, Stata, R, SAS and SPSS.

See **Annex 4: Calculation and use of survey weights** for more information on estimation method.

#### 5.1.2. Summarize the dataset

Once the weights and variables to describe the complex sample are constructed and the dataset is finalized, the dataset is summarized under the following categories.

1. Fieldwork and success of recruitment, including the number of:
  - clusters in survey and numbers of households found in visited clusters;
  - households visited and reasons for any not visited;
  - households agreeing, refusing, absent or non-response for other reasons;
  - eligible people surveyed, refusing, absent or non-response for other reasons;
  - people selected for specimen collection and the number of adequate specimens taken; and
  - eligible people unable to provide specimens, e.g. specimen collection was refused or specimen collection was unsuccessful.
2. Demographics of the survey sample, including information by stratum, sex and age group if applicable.
3. Biological specimen obtained. If a substantial portion of the specimens turned out to be unusable, describe how many and which strata and demographic groups they were from.
4. Laboratory data analysis, including the number of:
  - specimens analysed by each assay and, if appropriate, by laboratory
  - test runs with indications of invalid results requiring re-testing
  - valid positive, negative and equivocal results, and specimens re-tested
  - specimens with inadequate volume per team, per day and per region
  - specimens with repeated equivocal test results.

## 5. Analysis dataset using a codebook.

When the dataset has been assembled and cleaned, it is essential to generate a codebook (sometimes called a data dictionary) to help in the data analysis and interpretation. The data manager and statistician should review the codebook carefully to identify any remaining implausible values. An excellent codebook includes the following:

- **Overall summary:** A brief description of the study, sources of data, time period and manner in which data were collected, as well as the contact information for the client in case future codebook readers have detailed questions.
- **List of variables:** Simple uncluttered list of the variable names and labels for quick reference and electronic parsing.
- **Full dataset summary:** Summary of each variable in the dataset, documenting variable name, label, type and length. There should also be a summary of the variable in one of several fixed formats:
  - For categorical variables, a frequency table with data values, formatted labels and a count of the number and percent of observations in the dataset that take on that value.
  - For continuous variables, a univariate summary including minimum, maximum, median, mean, standard deviation, standard error, and the number of observations that are missing or that use special missing values (e.g. refused, don't know, questionnaire item skipped appropriately).
  - For dates, an indication of the first and last dates in the dataset.
- **Open-ended questions:** The codebook can either list the variable and the number of missing and non-missing responses, or it can document every unique verbatim answer in the dataset.
- **Stratum-specific summary:** Where there are well defined sub-groups, the responses from each sub-group may be documented separately. These data summaries are usually constructed, calculated and formatted using automated tools that can easily produce periodic updates to codebooks.
- **Notes:** Contains information about the dataset, including special documentation of data quality flags, problematic periods of data collection, formulae for calculating derived variables, known problems with individual variables, citations to literature that describes derived variables, and validated scales or scores calculated from raw survey responses.

### 5.1.3. Estimate prevalence

#### *Tabular Results*

After the dataset has been described and checked, estimates of seroprevalence and other population-level parameters are undertaken. Depending on the goals of the survey, the analysis plan may call for estimating population level totals, means or proportions.

The analysis software should account properly for the complex survey sample and incorporate the weights into the calculations. The estimation should use a set of commands saved in a programme file rather than in an interactive menu-driven session. Saving the analysis program will facilitate later modifications and independent review to reproduce and verify the results. Quantities described in the analysis plan are estimated, including seroprevalence for the antibody of interest in each stratum and, if appropriate, for all strata combined. Each estimated quantity will yield a point estimate and have a two-sided 95% confidence interval reported.

*(Note that the confidence interval for a proportion will be symmetric only when the point estimate is near 50% but will become more skewed as the point estimate approaches either 0% or 100%. A skewed asymmetric confidence interval is appropriate for an estimated proportion, with the longer side or tail of the distribution occurring on the side of the interval nearest 50%).*

If the results are being used to classify prevalence as likely to be above or below a fixed threshold of programmatic interest, then it is common to also calculate a one-sided 95% lower confidence bound (LCB) and a one-sided 95% upper confidence bound (UCB). The 95% LCB and UCB are not the same as the lower and upper bounds of the 95% confidence interval. The 95% LCB and UCB are calculated using the upper and lower bounds of a 90% confidence interval. These quantities are summarized in output tables. To maximize reproducibility, it is advisable to have the tables generated by a saved set of commands rather than have an analyst manually copy and paste results.

See World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual: (1) Section 6.5.

If seroprevalence was coded using only two codes (e.g. positive and negative), it is sufficient to summarize the estimated proportion of the population that are seropositive. The remainder (1-seropositive) are estimated to be seronegative. If the analysis dataset includes a third outcome code representing equivocal results, then it will be helpful to list the estimated population percent designated as negative and the percent equivocal, along with their confidence intervals and bounds. With all three categories tabulated explicitly, a reader will be able to assess whether it makes any difference if the equivocal results are truly positive (combining the equivocal and positive categories) or truly negative (combining the equivocal and negative categories).

### *Graphic results*

In addition to tables, it can be helpful to provide a graphical representation of estimated prevalence along with its confidence interval.

### *Formal statistical comparisons*

Some serosurveys will yield only descriptive results, where seroprevalence is documented for several antibodies and/or for several geographic or demographic strata. In most serosurveys, the analysis plan will include a formal comparison of population prevalence in different strata, e.g. comparing prevalence in children versus adults, males versus females, or different geographic regions.

See World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual (1) Section 6.4.

A hypothesis test can also be used to determine whether the data are statistically:

- higher than a fixed programmatic threshold, e.g. for herd immunity to measles or rubella
- lower than a fixed programmatic threshold
- higher than the prevalence measured in an earlier survey.

See World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual (1) Annex M and Annex N and Section 6: Use of serological data for modelling.

### *Identify clusters with surprisingly low seroprevalence*

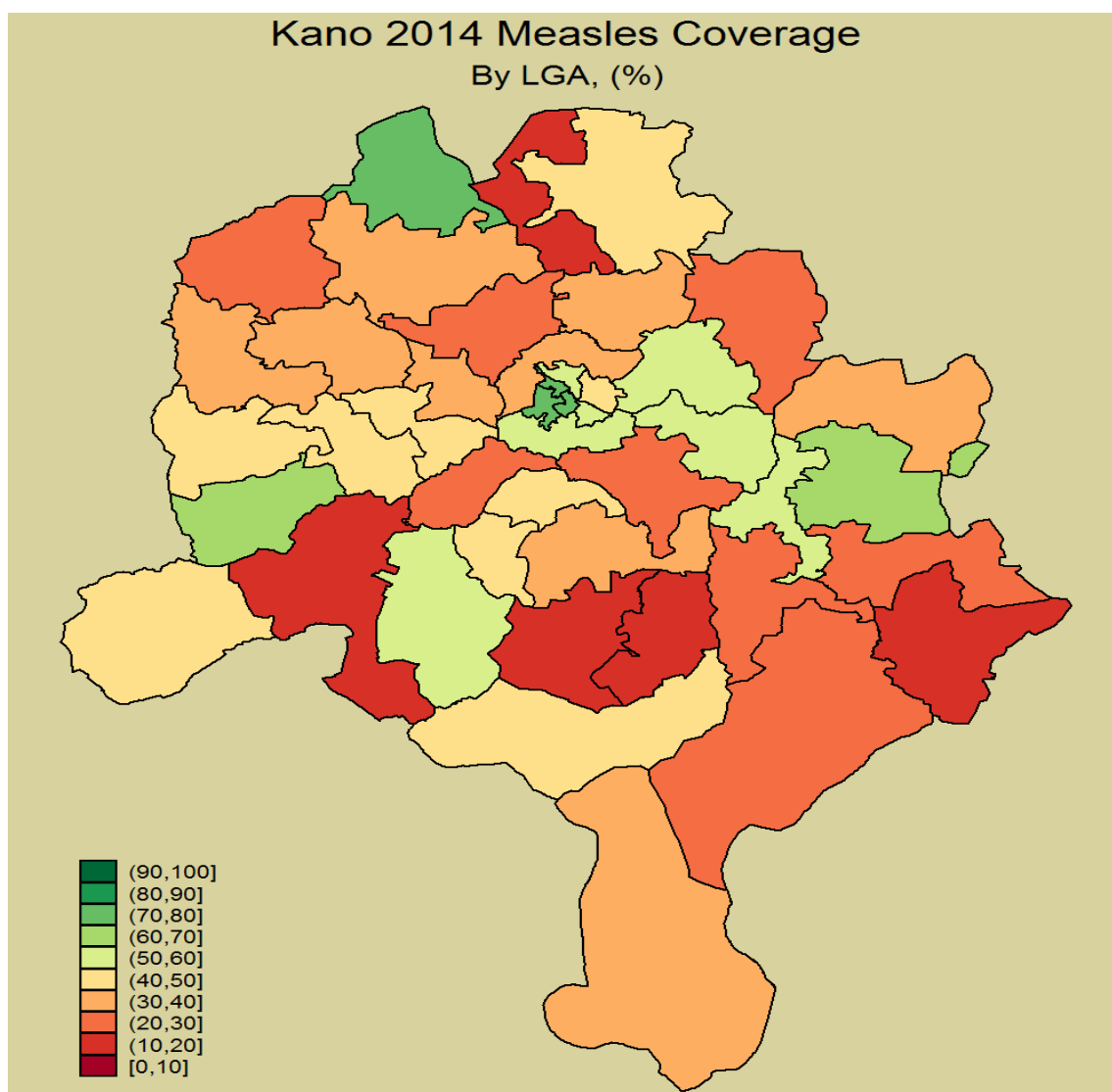
It may be appropriate to identify clusters where only a very small fraction of respondents is seropositive. Staff from the national immunization programme may wish to follow-up with inquiries to understand the reasons for localized susceptibility. Clusters with low coverage may be identified using a bar chart called an organ pipe plot (2).

See World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual (1) Section 6.1 for additional detail.

### Map Results

In addition to tabular results and graphical display of confidence intervals, it may sometimes be helpful to portray prevalence point estimates on a color-coded map (Figure 5-1). Maps with regions shaded to convey the outcome measure are called choropleth maps. They can be easily made using the statistical software packages listed above or geographic information system (GIS) software, such as ArcGIS ([www.esri.com](http://www.esri.com)) or Q-GIS, a free and open source GIS ([www.qgis.org](http://www.qgis.org)). Other types of data mapping are available. Where possible, data outside established lower limits of seroprevalence should be highlighted to initiate immediate investigation and action.

**Figure 5-1. Example of a choropleth map**



## 5.2. Interpret results

### 5.2.1. Results Interpretation

Even when a serosurvey is appropriately planned and implemented, the results must be interpreted with caution. It is known, for example, that a small proportion of individuals immunized against measles in infancy have low or undetectable antibody levels before a repeat vaccination (as with an SIA), and that antibody levels return to low levels several months after receipt of vaccine (3, 4). This is especially true in younger vaccine recipients. Therefore, measles seroprevalence data may be biased due to low or undetectable antibody levels in a proportion of vaccinated individuals.

In addition, the chance of classifying an immune person as susceptible may vary according to the assay used or the method employed to elucidate equivocal results. Various studies have suggested that antibody levels may wane over time if there is no boosting from further exposure to measles (5). It may be difficult, therefore, to determine the proportion of individuals without detectable antibody that are truly non-immune in settings where there has been little or no measles virus circulation for many years. For example, given knowledge of trends in coverage and incidence of children aged 10–15 years, a surprisingly high proportion of children may be found as seronegative, whereas the true percentage susceptible to infection is much lower due to waning antibody levels and the presence of cell-mediated immunity. This situation may arise because many individuals have antibody levels below the assay's cut-off but were in fact successfully immunized in infancy and are protected from disease by cell-mediated immunity. A rapid secondary antibody response following exposure to infection would be expected.

It is possible that the serosurvey results are unexpected and remedial actions are required. For example, if the seroprevalence is much lower than anticipated, the testing may be questioned. Negative samples may be referred for PRNT for confirmation. Further studies may be initiated:

- review of the amount of wild-type virus circulating in the population;
- assessment of the vaccine effectiveness in high endemic regions;
- determination if vaccinated individuals with low or negative antibody levels are protected; and
- investigation if an immune response due to existing cell-mediated immunity is developed on exposure (6).

## 5.3. Write report and share results

### 5.3.1. Writing the report

The primary objectives of a serosurvey report are to present the results in a way that is easily understood, to help the reader understand the strengths and weaknesses of the survey design, and to describe the implementation and execution of the protocol so readers can determine whether sampling and non-sampling errors are likely, and the potential impact of these errors.

Serosurvey reports should present a clear description of sampling methods, including their type (probability or nonprobability) and steps followed. A description of laboratory assays and testing strategies used and the criteria for determining immunity should be reported (e.g. cut-off values for seropositivity). Any limitations of the survey design and laboratory methods should be described and taken into account when interpreting serosurvey findings. Possible biases should be noted and addressed to the extent possible. An indication of how applicable the results are to other populations should also be discussed.

A template of a proposed report outline is included in **Annex 5**.

A primary reason for conducting serosurveys in support of measles and rubella elimination is to obtain additional supportive information or confirmatory evidence on progress made towards achieving the elimination goal. The primary question to be answered is whether the level of susceptibility found in the serosurvey is consistent with achieving and maintaining interruption of measles and rubella virus circulation. Any reporting of serosurvey results should attempt to answer this question or provide an explanation as to why the results cannot be used to answer this question.

In addition to the programmatic functions, serosurveys also have some political or administrative implications that need to be recognised and communicated with care as serosurvey results may:

- contradict other reported data for seroprevalence or vaccination coverage estimates, or indicate that immunization services are less effective than claimed;
- suggest that national authorities face population susceptibility problems that are beyond their resource capabilities, such as those that originate from, or have implications for transmission to, a neighbouring country where substantial, cross-border population movement occurs; and/or
- threaten to further stigmatize and isolate an already underserved and marginalized ethnic, economic or social group.

Serosurvey findings should be interpreted and communicated in light of current and historic data on the disease incidence, as well as the policies and performance of immunization programmes, including any past supplementary immunization activities (7). It is important to stress that seroprevalence and vaccine coverage measure different things; there are real biological reasons for discordance, especially if wild-type virus is circulating. Effective communication of result interpretation will help put into context the survey findings, which is particularly important for serosurveys based on convenience samples. Findings may highlight areas for improvement, such as higher seroprevalence in birth cohorts targeted by SIA than in birth cohorts relying only on routine immunization. This may suggest the need to strengthen routine immunization but also demonstrate that SIAs are a useful strategy in this setting (if natural immunity is ruled out as a reason for the age-differences). If other vaccine-preventable diseases have been included in the serosurvey, comparison of the results across different disease can help determine if the identified problems are related to a given vaccine or are of systemic nature related to the immunization programme in general.

### 5.3.2. Addressing potential bias

Reports should describe limitations in the survey and the methods employed to minimize biases. Authors should review the types of bias listed in **Section 2.7: Minimize survey error** and determine if they apply to the serosurvey. Brief comments on steps taken to mitigate the limitations and how remaining limitations might influence the results should be included in the report. It is considered best practice to list these sources of limitation clearly, to identify where they were successfully mitigated and to be honest about where they may influence the survey results. A clear description of the common problems with these surveys and measures taken to mitigate those problems may help persuade the reader that the survey results are indeed representative of the target population. Failures to comply with the survey protocol should be identified clearly and data that are of questionable quality for any reason should be flagged and should trigger a sensitivity analysis to indicate how the data quality issue(s) might affect the overall outcomes of interest.

### 5.3.3. Feedback and reports to stakeholders

The final results of the serosurvey should be shared with the local health authorities for appropriate dissemination to the community and to guide programmatic activities. The report may help in mobilizing civil groups to highlight and support the project to their communities. It is important to share information with all levels that participate in activities designed to reach elimination goals, to help with establishing and maintaining an effective disease control programme.

Consider providing access to survey results to policymakers, local EPI managers, health facility workers and senior health officials. The report's implications go beyond a national level to multiple administrative levels, such as provincial and districts levels. Communities covered by the serosurvey should also receive feedback, presented in ways appropriate to a lay audience.

For detailed discussion on sharing results see the World Health Organization Vaccination Coverage Cluster Surveys: Reference Manual (1) Section 7.7.

## References for Chapter 5

1. World Health Organization. 2018. Vaccination coverage cluster surveys: reference manual (WHO/IVB/18.09). World Health Organization, Geneva.
2. Prier M, Rhoda D. 2018. Organ Pipe Plots for clustered datasets - visualize disparities in cluster level coverage, abstr 2018 Stata Conference 41.
3. World Health Organization. 2008. The immunological basis for immunization series: module 11: Rubella. Biologicals. IVa, WHO, Geneva, Switzerland.
4. World Health Organization. 2009. The immunological basis for immunization series: module 7: measles - Update 2009. Immunization Vaccines and Biologicals, The immunological basis for immunization series : module 7: measles - Update 2009., Geneva.
5. Vyse AJ, Gay NJ, Hesketh LM, Pebody R, Morgan-Capner P, Miller E. 2006. Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in England and Wales at the beginning of the 21st century. *Epidemiol Infect* 134:1303-1312.
6. Picone O, Bouthry E, Bejaoui-Olhmann Y, Cordier AG, Nedellec S, Letourneau A, Carbonel M, Brollo M, Grangeot-Keros L, Ayoubi JM, Benachi A, Rouge E, Vauloup-Fellous C. 2019. Determination of rubella virus-specific humoral and cell-mediated immunity in pregnant women with negative or equivocal rubella-specific IgG in routine screening. *J Clin Virol* 112:27-33.
7. Winter AK, Martinez ME, Cutts FT, Moss WJ, Ferrari MJ, McKee A, Lessler J, Hayford K, Wallinga J, Metcalf CJE. 2018. Benefits and challenges in using seroprevalence data to inform models for measles and rubella elimination. *J Infect Dis* 218:355-364.