Supplementary S1 Text: Instructions for S1 Code

S1 Code: Generate Synthesis Data

Data Generating

The submitted dataset was generated using this file. Only the variables used in our analysis were generated. The column names and order followed the original dataset. In total, 4856 observations were included in the data and 14 variables were associated with our study. Note that although the process of generating synthesis data follows the general rules in the original data, the whole dynamic and underlying distributions may not be fully captured, so the results are different from what is demonstrated in the dissertation. The detailed explanation of the variables is in the *S1 Table*, and we simulated them using the following procedure:

- **ID-NIHE** A unique ID for each individual based on the examination year.
- **MAC-ELISA- Final result/ GAC-ELISA-Final result** A string represents the MAC-ELISA and GAC-ELISA test results for each individual based on the probabilities estimated in the original data.
- **Serotype** A string represents the PCR result with the respective probabilities approximate to the seroprevalence in the original data.
- **PRNT samples** A logical value based on the results of MAC-ELISA and GAC-ELISA. If both tested negative, then FALSE.
- **PRNT50** A string containing the serotypes with detectable antibody levels based on the PRNT50 DENV1-4 columns.
- **PRNT50 DENV1-4** Numerical values on a log2 scale for DENV1-4, with higher values for DENV1 and DENV2 followed the distribution in the original data.
- **Date of birth** A birth year generated based on the observed distribution in the original data. We treat the birth years later than the onset year as false-typed entries which will be removed in S2 Code to imitate the wrong entries in real data.

- **Gender** Equal probability of gender assigned.
- **Date of onset/ Examination date** Various formats that occurred in the original data were sampled randomly to reflect the data variability and the years are constrained to 2020-2022 which is our study period.

Initial Data Cleaning

Six more variables were generated based on the given variables for further analysis.

- **OnsetDate** The cleaned format of the date of onset.
- **OnsetYear/ OnsetMonth** The onset year and month of extracted from the "ParsedDate".
- **BirthDate** The cleaned format of the date of birth.
- **BirthYear** The birth year extracted from date of birth.
- **Age_clean** The age of the patient is calculated from the onset year and birth year.