

**Official study title:**

Whole Genome-Based Surveillance of Nasopharyngeal Pneumococcal Carriage and Serotype Distribution Among Vietnamese Children Hospitalized With Acute Respiratory Symptoms

**ClinicalTrials.gov Identifier (NCT Number):**

NCTXXXXXXXX (to be added once assigned)

**Organization's Unique Protocol ID:**

VNCH-TRICH-2026-50

**Sponsor:**

Vietnam National Children's Hospital (VNCH), Hanoi, Vietnam

**Responsible Party:**

Principal Investigator, Vietnam National Children's Hospital

**Document Type:**

Study Protocol and Statistical Analysis Plan

**Document Version:**

Version 2.0

**Document Date**

15.02.2026

# STUDY PROTOCOL

## 1. Study Title

Whole Genome-Based Surveillance of Nasopharyngeal Pneumococcal Carriage and Serotype Distribution Among Vietnamese Children Hospitalized With Acute Respiratory Symptoms

## 2. Background and Rationale

Nasopharyngeal carriage of *Streptococcus pneumoniae* is a prerequisite for invasive pneumococcal disease and serves as an early indicator of vaccine impact and serotype replacement. In Vietnam, PCV10 and PCV13 have been introduced in the private sector; however, updated post-vaccine serotype data are lacking, particularly using whole genome sequencing (WGS).

Conventional PCR-based serotyping has limitations in detecting co-colonization and non-typeable strains. WGS enables high-resolution serotyping, antimicrobial resistance (AMR) gene detection, and lineage analysis. This study aims to generate contemporary genomic surveillance data to inform national pneumococcal vaccination strategy.

## 3. Objectives

Primary Objective

To determine:

- Prevalence of nasopharyngeal pneumococcal carriage
- Serotype distribution identified by whole genome sequencing

Among children aged 2–59 months hospitalized with acute respiratory symptoms.

Secondary Objectives

1. Compare carriage prevalence and serotype distribution between:
  - Previously healthy children
  - Children with chronic underlying conditions
2. Identify predictors of carriage (age, vaccination history, antibiotic use, daycare attendance, environmental exposures).
3. Detect:
  - Co-colonization patterns
  - Antimicrobial resistance genes

4. Estimate theoretical vaccine coverage of circulating serotypes (PCV10, PCV13, PCV15, PCV20).

#### **4. Study Design**

- Design: Hospital-based cross-sectional study
- Setting: Vietnam National Children's Hospital (VNCH), Hanoi
- Sample size: 530 children (265 per group)
- Sampling method: Convenience sampling during working hours

#### **5. Study Population**

##### Inclusion Criteria

- Age 2–59 months
- Hospitalized with acute respiratory symptoms ( $\leq 14$  days)
- Informed consent from parent/guardian

##### Exclusion Criteria

- Confirmed pneumococcal infection within 30 days

##### Study Groups

- Group 1: Previously healthy
- Group 2: Chronic underlying conditions

#### **6. Procedures**

- Structured clinical and demographic questionnaire
- Nasopharyngeal swab collection (WHO protocol)
- Culture and MALDI-TOF identification
- Whole genome sequencing (Illumina platform)
- Bioinformatics analysis:
  - In silico serotyping (PneumoCaT)
  - MLST
  - AMR gene detection

- Variant analysis

## **7. Outcomes**

### Primary Outcomes

1. Pneumococcal carriage prevalence
2. Serotype frequency distribution

### Secondary Outcomes

- Between-group difference in carriage prevalence
- Adjusted predictors of carriage
- Co-colonization rate
- Prevalence of AMR genes
- Estimated vaccine-type (VT) coverage

## STATISTICAL ANALYSIS PLAN

### 1. General Principles

- Statistical significance: two-sided  $\alpha = 0.05$
- 95% confidence intervals (CI) reported
- Software: STATA version 17
- All analyses predefined prior to database lock

### 2. Sample Size Justification

Based on expected carriage prevalence of 45%, precision 6%, 95% confidence level:

$n \approx 265$  per group

Total sample size: 530

This sample size provides:

- Adequate precision for prevalence estimation
- Sufficient power (>80%) to detect ~12–15% difference between groups

### 3. Analysis Populations

Full Analysis Population (FAP)

All enrolled participants with valid nasopharyngeal sample results.

Microbiologically Confirmed Population

Participants with confirmed *S. pneumoniae* detection.

### 4. Descriptive Analysis

Baseline Characteristics

Continuous variables:

- Mean  $\pm$  SD (normal distribution)
- Median (IQR) (non-normal distribution)

Categorical variables:

- Frequencies and percentages

Variables include:

- Age
- Sex
- Birth weight
- PCV vaccination status
- Antibiotic use
- Daycare attendance
- Household smoking
- Comorbidities

## 5. Primary Outcome Analysis

### 5.1 Carriage Prevalence

Prevalence (%) =  
 (Number positive for *S. pneumoniae* / Total enrolled)  $\times$  100

Reported with:

- 95% CI using exact binomial method

Group comparison:

- Chi-square test
- Fisher's exact test if expected cell count <5

### 5.2 Serotype Distribution

- Frequency distribution of serotypes
- Relative proportion (%)
- Diversity indices (optional, if included):
  - Simpson's index
  - Shannon diversity index

Between-group comparison:

- Chi-square test for distribution differences

## 6. Secondary Analyses

### 6.1 Predictors of Pneumococcal Carriage

### Step 1: Univariable Analysis

Logistic regression for each variable:

Outcome:

Carriage status (Yes/No)

Predictors:

- Age (continuous and categorical)
- Sex
- Vaccination status
- Antibiotic use
- Daycare attendance
- Household size
- Smoking exposure
- Underlying condition

Variables with  $p < 0.20$  will be entered into multivariable model.

### Step 2: Multivariable Logistic Regression

Model:

$$\text{logit}(P(\text{carriage})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Outputs:

- Adjusted Odds Ratios (aOR)
- 95% CI
- p-values

Model diagnostics:

- Hosmer–Lemeshow goodness-of-fit
- Variance Inflation Factor (VIF) for multicollinearity

## 6.2 Co-Colonization Analysis

Outcome:

Presence of  $\geq 2$  serotypes

Analysis:

- Proportion with 95% CI
- Comparison between groups (Chi-square)

Optional:

Multivariable logistic regression for predictors of co-colonization.

### 6.3 Antimicrobial Resistance Gene Analysis

- Frequency of each AMR gene
- Proportion resistant by genotype
- Comparison between groups

Optional:

Association between serotype and AMR gene presence  
(Logistic or multinomial regression)

### 6.4 Vaccine Coverage Estimation

Each identified serotype mapped to:

- PCV10
- PCV13
- PCV15
- PCV20

Coverage (%) calculated as:

$(\text{Number of isolates covered by vaccine} / \text{Total isolates}) \times 100$

Between-group comparison via Chi-square test.

## 7. Subgroup Analyses

Pre-specified subgroups:

- Age <12 months vs  $\geq 12$  months
- Fully vaccinated vs partially/unvaccinated
- Recent antibiotic use (Yes/No)

Interaction terms tested in regression models.

## 8. Missing Data Handling

- Proportion of missing data reported

- If missing >10% for key covariates:
  - Multiple imputation (MICE method)
- Sensitivity analysis:
  - Complete case vs imputed model

#### 9. Sensitivity Analyses

1. Excluding children with recent antibiotic use
2. Restricting analysis to culture-confirmed isolates
3. Reanalysis using alternative age categorization

#### 10. Data Presentation

Results will be presented as:

- Tables (baseline, regression outputs, serotype distribution)
- Forest plots (adjusted ORs)
- Bar charts (serotype frequency)
- Pie charts (vaccine coverage distribution)