Statistical Analysis Plan (SAP)

The Fever without Source Study

SAP version: **1.3**
File created: **28.12.2015**
Last changes: **28.03.2017**
Author: **Dr Ralf Krumkamp**

Table of contents:

1. Objectives 2
2. Study group, study area 2
3. Study design 2
4. Study outcomes 2
5. Sample size 3
6. Descriptive analysis 3
7. Measures of association 3
8. Missing data 4
9. Data management 4

1. Objectives

The Fever without Source Study (FWS) was designed to systematically assess clinical (i.e., based on disease symptoms) and microbiological (i.e., based on identified microbiological pathogens) diagnoses in children admitted to a hospital in a rural region in Ghana. Aims of the study were

- an exact diagnosis and identification of causes of severe febrile illnesses in hospitalised children,
- the determination of the frequency and distribution of severe febrile illnesses in hospitalised children,
- to link microbiological and clinical diagnoses, and
- the analysis of co-infection dynamics with the focus on *Plasmodium falciparum* parasitaemia.

2. Study group, study area

Children (≥30days and ≤15 years) with fever (≥38°C) admitted to the Children's Ward of the Agogo Presbyterian Hospital (APH) in the Asante Akim North Municipality, Ashanti Region, Ghana, were eligible to participate in the study. Furthermore, children below 15 years of age with a temperature <37.5°C and without signs of infection were recruited at vaccination clinics in the surroundings of the APH. Participants were recruited between November 4th, 2013 and April 30th, 2015

Asante Akim North has an estimated population of 142,400 inhabitants, spread over an area of 1,160 square kilometres. The region has a tropical climate and is mainly covered by secondary rain forest and cultivated land. Malaria is highly endemic in that area with seasonal peaks.

3. Study design

At admission, children were recruited in the study and clinical and laboratory parameter were determined. Hence, the study can be classified as cross-sectional conducted over a period of 18 months.

4. Study outcomes

Primary study outcomes are established clinical diagnoses based on observed disease symptoms and additional clinical parameters. Main diagnoses considered are *Plasmodium* spp. infection, lower respiratory tract infection (LRTI) and pneumonia, urinary tract infection (UTI), gastrointestinal and abdominal infection, bloodstream infection, and infection of the

central nervous system or meningitis. The remaining rare diagnoses are summarised into the category "others".

Within these diagnoses, the presence of potentially causative pathogens are analysed, which builds the respective subgroups of diagnoses with and diagnoses without pathogens detected.

5. Sample size

The overall sample size is calculated to estimate the proportion of a diagnosis with sufficient precision. Considering a prevalence of a diagnosis of 8% within the study group, a sample size of 984 would be needed to retrieve an estimate with 5% precision (95%-confidence interval [CI]). Hence, a sample size of 1,000 was considered as the minimum sample size for the study. ‘

6. Descriptive analysis

Descriptive statistics are applied to summarise patient data. Categorical variables are displayed as frequencies with percentages, non-normally distributed continuous data as the median along with the interquartile range (IQR) and normally distributed continuous variables as the mean with the standard deviation (SD).

Estimating the frequency of diagnoses in study children is a primary study outcome, which is analysed using single proportions. Subgroups are considered in order to assess whether diagnoses differ among strata of patient characteristics. To display age effects, the four age categories 0–<1, 1–<2, 2–<4 and ≥4 years as well as the dichotomous indicator with categories ≤2 and >2 are established. Based on individual parasite counts four categories are introduced, namely (i) no malaria parasites detected, (ii) 1–10,000 parasites per μl, (iii) >10,000 to 100,000 parasites per μl, and (iv) patients with above 100,000 parasites per μl.

7. Measures of association

To evaluate the likelihood of a co-infection in children with parasitaemia the risk ratio (RR) with corresponding 95%-confidence interval (CI) is calculated. The RR is the ratio of the fraction of children with a diagnosis of interest in all children with parasitaemia to the fraction of children with this diagnosis but without parasitaemia. In the present data associations between diagnoses are likely to be confounded by age. Hence age-stratified analyses using the Mantel-Haenszel method is considered. Parasitaemia is defined as presence of fever along with any *Plasmodium* spp. parasite in a blood sample.

Due to the exploratory nature of the data analysis, no significance testing is applied. All data analyses are performed with Stata 14 (StataCorp LP, College Station, USA).

8. Missing data

The study group used for the analysis is selected to have full data in the core variables, namely basic patient characteristics (e.g., age and sex) and established clinical diagnoses and in the microbiological diagnoses malaria and bacteraemia. Pathogens were analysed according to a clinical algorithm. For example, stool samples were taken in children with diarrhoea to test for the presence of gastrointestinal pathogens. Due to lacking samples, in some cases laboratory results can be missing, which were excluded from the analysis. Thus, the distribution of pathogens is based on the collected patient samples. Since we assume that samples are missing completely at random (i.e., the pathogen present in a sample does not affect the likelihood to miss out a sample), we believe that the proportion of pathogens are not biased due to missing values. However, the overall frequency will be underestimated.

Systematic missings (e.g., in HIV diagnosis due to withdrawn informed consent) will be described and the actual denominator will be presented in the respective calculations.

9. Data management

The 4th Dimension database management system is used to enter and manage the study data. The database implements plausibility checks in order to highlight implausible entries during the entry process. Data is double entered and in case of inconsistency entries are evaluated and revised. The data is pseudonymised, thus throughout data entry, data management and data analysis no link between an observation and an individual can be established.