Statistical analysis plan for the 04EN study (ISRCTN88101063)
"An evaluation of the effect of probiotics in children with acute watery diarrhoea attending Children's Hospital 2 in Ho Chi Minh City, Vietnam"

Authors: Nhat Le Thanh Hoang, Anh Nguyen Duc, Chau Tran Thi Hong, Stephen Baker, Marcel Wolbers
Version: Final version before unblinding of the trial, 17Dec2015

Purpose

This document details the planned analyses and endpoint derivations for the ISRCTN88101063 trial as outlined in the study protocol version 6.2. Of note, an early version of the study protocol has been published (Kolader et al., Trials 2013, 14:27, doi:10.1186/1745-6215-14-27) but some endpoint definitions have been changed prior to starting the study. In particular, for reliability and practicality reasons, the primary endpoint is assessed by the participant's parent/guardian according to the final protocol rather than the nurse on duty (as per the Trials publications) leading to a right-censored rather than a interval-censored primary endpoint.

This analysis plan focuses on the analysis for the main clinical trial publication and does not include analysis for any subsidiary studies.

Statistical software

Data derivations will be performed with the statistical software SAS v9.4 (SAS Institute, Cary, North Carolina, US). All statistical analyses will be performed with the statistical software R using the current R version at the time of the final analysis (R Foundation for Statistical Computing, Vienna, Austria).

Analysis populations

Intention-to-treat population (ITT)

The main analysis population for all analyses is the full analysis set including all randomized patients and analysis is according to the randomized treatment arm.

Per-protocol population

The main comparison of the primary endpoint will also be performed in the per-protocol population which excludes the following participants:

- Patients who violated an inclusion or exclusion criteria, i.e. those with (2='no') for any inclusion or (1='yes') to any exclusion criterion.
- Patients who received less than 8 doses of study treatment (excluding attempted but not taken doses due to vomiting, spilling or other reasons).
- Patients who used probiotics by themselves (i.e. in addition to randomized treatment), i.e. those with [CONMED.OTHDRUG1NAME= "LACBIO" or "PROBIO",…, CONMED.OTHDRUG4NAME= "LACBIO" or "PROBIO"].
- Patients who were lost to follow-up or withdrew from the study, i.e. those with [FOLLOWUP.ASSESSND=1 or 2].

Baseline characteristics

The baseline date and time is defined as the date and time the first dose of study drug is given or attempted to be given (DRUG.DATEGIVEN and DRUG.DRUGTIM with DRUG.DOSE=1). If no drug was given, the date and time of enrolment will be used instead (BASE.ENDATE and BASE.ENROLTIM).

Baseline characteristics will be summarized as median (IQR) for continuous data and n (%) for categorical data. The amount of missing data for each baseline characteristic will also be displayed.

Formal comparisons of baseline characteristics between study arms are discouraged by most statisticians (see e.g. Senn SS (2008): Statistical issues in Drug Development, 2nd Edition, Wiley [p. 98f]) but mandated by some journals. To satisfy all potential publishers, we will calculate p-values (based on the Wilcoxon rank sum test and Fisher's exact test for continuous and categorical data, respectively) but will only report them if mandated by the journal.

The following baseline characteristics will be summarized by treatment arm [with derivation rules in brackets]:

DEMOGRAPHICS, HISTORY AND EXAMINATION

- **Demographic data:**
  a. Sex (male/female) [BASE.SEX; 1= "female", 2= "male"]
  b. Age (in months)
     [DATEOFBIRTH=date created from BASE.DOBM/BASE.DOBD/BASE.DOBY]
     [AGE= (Baseline date-DATEOFBIRTH+1)/30.4]
- **General examination at enrolment (Day 1, 1st assessment)**
  a. Temperature (DAILYD1.TEMPEXAM)
  b. Pulse (DAILYD1.PULSEEXAM)
  c. Weight (DAILYD1.WEIGHT)
- **Symptoms occurring during this illness episode:**
  a. Duration of diarrhoea at baseline (in hours): Baseline date / time – BASE.DATEADIARR/BASE. ONSETTIM
  b. Aproximate average number of diarrhoeal episodes per day (times) [BASE.DIARRTIMES]
  c. Vomitting (yes/no) [BASE.VOMIT; 1= "yes", 2= "no", 3= "unknown"]
  d. Vomitting times [BASE.VOMITTIMES]
  e. Fever (yes/no) [BASE.FEVER, 1= "yes", 2= "no", 3= "unknown"]
  f. Lethargic (yes/no) [BASE.LETHAGIC, 1= "yes", 2= "no", 3= "unknown"]
  g. Abdominal pain (yes/no) [BASE.ABDOMINALP, 1= "yes", 2= "no", 3= "unknown"]
  h. Anorexia (yes/no)= [BASE.ANOREXIA, 1= "yes", 2= "no", 3= "unknown"]
- **Enrolment assessment of diarrhea and vomiting (Day 1, 1st assessment)**
  a. Diarrhea in last 12 hours? (DAILYD1. DIARRL12H)
  b. Number of diarrhea episodes in last 12 hours (DAILYD1.DIARRTIME12H)
  c. Aspect of diarrhea ("Mucoid" if DAILYD1.DIARRMUCOID=1, "Bloody" if DIARRBLOODY=1, "Unknown" if DIARRNOTSEE=1, "Watery/watery with solids" if DIARRWATERY=1 and DIARRWATSOLID=1, "Watery" if only DIARRWATERY=1, "Watery with solids" if only DIARRWATSOLID=1)
  d. Vomiting in the last 12 hours (DAILYD1. VOMITL12H)
  e. Number of vomiting episodes in last 12 hours (DAILYD1.TIMESVOML12H)
  f. Yoghurt last 12 hours (DAILYD1.YOGHURT)
- **Previous history:**

a. Number of diarrhoea epidosed in the past year (times) [BASE.DIARRTIMEPYEAR]
b. Admission to the hospital for diarrhoea in the past year [BASE.ADMITDIARR, 1= “yes”, 2= “no”, 3= “unknown”]
c. Treatment with antimicrobials in the past month [BASE.ANTMICROBIALPM, 1= “yes”, 2= “no”, 3= “unknown”]
d. Treatment with probiotics in the past week (prior treatment with probiotics) [BASE.PROBIOTREAT, 1= “yes”, 2= “no”, 3= “unknown”]
e. Treated with anti-diarrhoeal medication in this disease period [BASE.DIARRTREATDP, 1= “yes”, 2= “no”, 3= “unknown”]
f. Treated with anti-emetic medication this disease period [BASE.ANTMEEMETIC, 1= “yes”, 2= “no”, 3= “unknown”]
g. Treated with antimicrobials in this disease period [BASE.ANTIMICROBIO, 1= “yes”, 2= “no”, 3= “unknown”]

- **Intake of food during this disease episode:**
  a. Child has been eating normally [BASE.NINTAKEFOOD, 1= “yes”, 2= “no”, 3= “unknown”]
  b. Child had yoghurt [BASE.YOGHURTDE, 1= “yes”, 2= “no”, 3= “unknown”]
  c. Soybeans or soybean products eaten [BASE.SOYBEANDE, 1= “yes”, 2= “no”, 3= “unknown”]
  d. Child currently breastfeeding [BASE.BREASTFEDD, 1= “yes”, 2= “no”, 3= “unknown”]
  e. Child drinking formula milk [BASE.FORMULARMILK, 1= “yes”, 2= “no”, 3= “unknown”]

STOOL MICROLOGY RESULT

- Leucocytes [LABMICRST.LEUCOCYTES, 1= “yes”, 2= “no”]
- Erythrocytes [LABMICRST. ERYTHROCYTES, 1= “yes”, 2= “no”]
- Parasites [LABMICRST.PARASITES, 1= “yes”, 2= “no”]
- Parasite types [LABMICRST. PARASITESSPE1] and [LABMICRST. PARASITESSPE2]
- Rotavirus [LABMICRST. ROTAVIRUS, 1= “yes”, 2= “no”]
- Rotavirus viral load at baseline [STOOL_PCR.ROTAVIRUS_VIRALLOAD=1, with STOOL_PCR.STDDAY=1]
- Norovirus [LABMICRST.NOROVIRUS, 1= “yes”, 2= “no”]
- Norovirus viral load at baseline [STOOL_PCR.NOROVIRUS_VIRALLOAD=1, with STOOL_PCR.STDDAY=1]
- Campylobacter [LABMICRST. CAMPYLOBACTER, 1= “yes”, 2= “no”]
- Campylobacter type [LABMICRST. CAMPYLOBACISO, 1= “C.Coli”, 2= “C.jejuni”, 3= “Campylobacter spp”]
- Shigella [LABMICRST.SHIGELLA, 1= “yes”, 2= “no”]
- Salmonella [LABMICRST.SALMONELLA, 1= “yes”, 2= “no”]
- Other bacteria [LABMICRST.BECTERIA, 1= “yes”, 2= “no”]

HEMATOLOGY AND CHEMISTRY

- Hematocrit (HCT %)=[LAB_HEMA.HEMATOCRIT]
- White blood cell count (K/uL)=[LAB_HEMA.WBC]
  a. Neutrophils (%)=[LAB_HEMA.NEUTROPHILS]
  b. Lymphocytes (%)=[LAB_HEMA.LYMPHOCYTES]
  c. Eosinophils (%) =[LAB_HEMA.EOSINOPHILS]
- Platelet (K/UI)=[LAB_HEMA.PLT]
- Sodium (Na^+) (meq/l)=[LAB_CHEMIS.SODIUM]
- Potassium (K^+) (meq/l)=[LAB_CHEMIS.POTASSIUMKP]

- Urea (g/l)=[LAB_CHEMIS.UREA]
- Creatinine (mg/l)=[LAB_CHEMIS.CREATININE]

Planned analyses:
Baseline table for all variables as detailed above for the ITT population stratified by treatment group.

Primary endpoint – the time from the first dose of study medication to the start of the first 24 hour diarrhoea-free period

Derivation
Twice daily diarrhoea assessments (AM and PM) are recorded on the CRF DAY1 (dataset DAILYD1) and CRF DAY2-6 (dataset DAILY) forms and all assessments (in-patient, out-patient, by phone, or extracted from patient diary) will be used for the derivation. For the derivation, the two datasets DAILYD1 and DAILY will be combined and transformed to long format with one row per assessment, i.e. two rows per day if both AM and PM assessments were performed.
The start of the first 24 hour diarrhoea-free period is defined as the date and time of the last diarrhoeal episode after baseline (LASTDIARRTIM for AM assessments, LASTEPTIM for PM assessments) which is followed by a documented period of ≥24 hours without diarrhea. (If the documented period is <24 hours long but contains at least two subsequent assessments without diarrheal episodes and no further diarrhea assessments after that, this will also be accepted.)
For patients who have no diarrhoea recorded during the first 24 hours after baseline, the start of the last diarrhoeal episode will be defined as baseline if they have at least two documented subsequent assessments without diarrhoea. Otherwise, they will be treated as right-censored at baseline.
Patients without a 24 hour diarrhoaea-free period (as defined above) will be right-censored at their last recorded time of a diarrhoeal episode to indicate that their diarrhoaea was still ongoing or had not ceased for 24h at their last daily follow-up visit.

The time to the primary endpoint (in hours) is then defined as:
(datetime of last episode of diarrhea or censoring) – (datetime of baseline) + 1
The corresponding event indicator is 1 if cessation of diarrhoea was observed and 0 for censored patients.

Planned analyses
- The primary endpoint will be compared between the study arms on the basis of a *lognormal accelerated failure time* (AFT) regression model, which models the log-transformed outcome as depending linearly on covariates plus a normally distributed error term for unexplained variation. This method generalizes the standard linear regression model of the log-transformed outcome to censored outcome data as required for analyzing the primary endpoint and is implemented in R package “survreg”.
- The primary analysis assumes an approximate normal distribution for the log-transformed primary endpoint which may not reflect the truth. Thus, the following additional sensitivity analyses will be performed:
    a. An alternative parametric Weibull AFT models with treatment as the only covariate
    b. A Cox regression analysis with treatment as the only covariate
    c. A lognormal AFT as specified above but excluding subjects without any diarrhoea recorded during the first 24 hours after baseline (i.e. those with diarrhea clearance at baseline)

d. Comparison of different parametric AFT models (log-normal, Weibull, log-logistic, and generalized gamma) in terms of the Akaikie information criterion (AIC) and graphical checks

In case these sensitivity analyses provide clear evidence that the pre-defined lognormal model is mis-specified and the Weibull and the Cox model, i.e. sensitivity analysis a. and b., are consistent with each other but lead to discrepant conclusions from the lognormal model whether treatment is significant or not, the Weibull AFT model will replace the lognormal model as the primary analysis.

- Homogeneity of the treatment effect will also be assessed according to the following *pre-defined subgroups*:
  a. Age (stratified as 9-12, 13-24, 25-36, 36-48, and 49-60 months)
  b. Prior treatment with antibiotics: Yes/No.
  c. Prior treatment with probiotics: Yes/No.
  d. Pathogen: rotavirus (including any co-infections), norovirus (including any co-infections), rotavirus&norovirus (including any co-infections), any other pathogen, no pathogen identified

  Potential heterogeneity of the treatment effect across sub-groups will be tested using likelihood ratio tests for an interaction term between treatment and the grouping variable.
- Kaplan-Meier curves by treatment arm will be displayed and visually compared to the fits from the parametric AFT models.
- Regression AFT modelling for the primary endpoint including the following covariates (in addition to the treatment arm): duration of diarrhoea prior to enrolment, prior treatment with antibiotics (yes/no), prior treatment with probiotics (yes/no), age, and pathogen (3 indicator variables: rotavirus yes/no, norovirus yes/no, other pathogen yes/no).
- Other exploratory analysis will be performed as appropriate.

Secondary endpoint – The total duration of diarrhoea

Total duration = Duration of diarrhoea at baseline + duration after baseline (i.e. the primary endpoint). Planned analyses are the same as for the primary endpoint.

Secondary endpoint – Treatment failure

Derivation:

Treatment failure is defined as the occurrence of any of the following:

- no resolution of diarrhoea during the 5 day treatment course (i.e. primary endpoint reached or censored at >120 hours)
- severe symptoms for which treatment is stopped (any serious adverse event leading to stopping of study drugs, i.e. any record in dataset SAE with SAESTDDRUG=2 ('Stopped'))
- requirement for additional anti-diarrhoeal treatment (a record in dataset CONMED with anti-diarrhoal= 'Yes', i.e. ATDIARR=1].

Indicators for overall treatment failure, the specific cause of failure plus the name of the anti-diarrhoeal drug (ATDIARRSMECTA=1, HIDRASEC=1, or ATDIARROTH ('Other')) will be derived.

Planned analyses:

The treatment failure will be compared between the two arms with a logistic regression model with treatment as the only covariate. Patients lost to follow-up without prior document treatment failure will be considered to not have experienced treatment failure.

Secondary endpoint – Stool frequency in the first 3 days

Derivation:

The total number of diarrhoea episodes or diapers during the first 3 days (72 hours) after the first dose of study drug will be derived based on twice daily diarrhea counts recorded in datasets DAILYD1 and DAILY. In addition, the individual stool frequencies during days 1, 2, and 3 will also be derived.

Planned analyses:

The stool frequency in the first three days will be compared between the two arms based on a quasi-Poisson regression models with the treatment arm as the only covariate. Quasi-likelihood, i.e. "family=quasipoisson()" in R function "glm", will be used to account for potential overdispersion.

Secondary endpoint – The norovirus and rotavirus viral load

The norovirus and rotavirus viral load in copies per millilitre (ml) of viral transport medium, by PCR on faecal swabs will be assessed daily from enrolment until discharge and at the patient follow-up visits. As these lab data are not part of the clinical database, they will be recorded by the lab technician in excel/csv format and transferred to the statistician.

Planned analyses:

Analysis of noro- or rota-virus viral loads, respectively, will be only for patients with a positive viral load for the respective pathogen at enrolment. Patient profiles of log-transformed longitudinal viral load measurements will be graphically displayed by randomized treatment arm. The main summary measure is the viremia AUC, defined as the area under the log-transformed viremia curve between enrolment (day 1) and day 7 obtained via linear interpolation of measured viral load measurements. Viral load measurements below the detection limit (Rotavirus: 500 [c/ml] and Norovirus: 5 [c/ml]) will be assigned half the detection limit for the calculation of the viremia AUC. Comparison between the two treatment arms will be based on linear regression with the treatment arm as the main covariate and adjustment for the log-transformed baseline viral load to increase power. In addition to the comparison of the viremia AUC, we will also fit longitudinal linear mixed effect models to the log-viremia data with a fixed intercept, fixed treatment arm specific slopes, and random (patient-specific) incercepts and slopes.

Secondary endpoint - The extent of intestinal L. acidophilus colonization

The extent of intestinal L. acidophilus colonization will be assessed by analyses of stool samples collected at enrolment, on discharge/or last day of daily follow up (one day after finishing the treatment course), and at outpatient follow-up visit, seven days after finishing the treatment course for the first 50 agreeing parents/guardians of participants who are assessed at the follow-up visit.

Intestinal L. acidophilus colonization will be measured at the later time point and the analysis strategy described elsewhere.

Secondary endpoint - The duration of hospitalization

Derivation:

[Date/time of hospital discharge]-[Date/time of hospital admission]+1 (in hours)

The date/time of hospital is recorded in BASE.DATEADMIT/BASE.ADMITTIM, and the date of discharge in OUTCOME.DISDATE/OUTCOME.DISTIM.

Planned analyses:

The duration of hospitalization will analyzed in the same way as the primary endpoint.

Serious adverse events

Serious adverse events are collected in dataset SAE. The frequency of serious adverse events will be summarized by treatment arm and the number of patients with events will be compared between the two groups based on Fisher's exact test.

Additional pre-defined analyses

- Recurrent diarrhea, i.e. a new diarrea episode since the initial episode as assessed at the day 13 follow-up visit (recorded in variable FOLLOWUP.NEWEPISODE, 1= "yes", 2= "no", 3= "initial episode has not stopped"]. Recurrent diarrhea will be analysed in the same way as the secondary endpoint treatment failure.
- Vomitting frequency in the first 3 days: this will be derived and analyzed in the same way as the secondary endpoint stool frequency.