Statistical Analysis Plan (SAP)

SAP version number: 2023–12–11

Person writing the SAP: Paul Blanche

Principal investigators: Allan Bybeck Nielsen & Ulrikka Nygaard

Protocol title: Oral-only Antibiotics for Bone and Joint Infections in Children

ClinicalTrials.gov ID: NCT04563325

Note: inspiration to write this SAP came from the SAP template provided by **TransCelerate** [1] as well as recommendations from Gamble et al. [6], Stevens et al. [14] and Evans and Ting [3].

[1] Available via: https://www.transceleratebiopharmainc.com/assets/clinical-content-reuse-solutions/

Contents

1 Statistical analysis plan approval signature page

Person writing the SAP:

Name: Paul Blanche

Position/Title: Associate Professor in Biostatistics

[redacted]

December 11, 2023

Signature & Date

Approved by:

Name: Allan Bybeck Nielsen

Position/Title: Senior registrar, MD

[redacted]

December 11, 2023

Signature & Date

Name: Ulrikka Nygaard

Position/Title: Consultant, associate professor, Ph.D.

[redacted]

December 11, 2023

Signature & Date

2 List of Abbreviations

- BJI: Bone and Joint Infection
- CI: Confidence interval
- CRP: C-reactive protein
- IVT: IntraVenous Treatment
- OT: Oral Treatment

3 Introduction

This document is the statistical analysis plan (SAP) for a randomized (1:1), non inferiority, clinical trial about initial high-dose oral antibiotics treatment (OT) versus initial intravenous antibiotics treatment (IVT) among children with bone and joint infection (BJI). The primary outcome is the presence of sequelae at 6 months. The protocol of this study was previously published, by Nielsen et al. [11]. There are no major changes to the protocol-planned analyses, but a few minor changes exist. These are summarized in Section 6.5.

3.1 Objectives, Endpoints, and Estimands

3.1.1 Primary

The primary objective is to show that initial high-dose Oral Treatment (OT) is not inferior to initial IntraVenous Treatment (IVT) with respect to the risk of sequelae after 6 months. A non inferiority margin of 5% difference in risk was pre-specified. Accordingly, the primary outcome is the presence of abnormal mobility or function of the affected joint/bone at 6 months follow-up, as evaluated by a blinded clinical examination by a qualified clinician.

Specifically, the two interventions being investigated are initiating either OT or IVT and continuing the treatment until the first of these two events occurs:

- clinical improvement (pain and mobility) and paraclinical improvement (decreasing C-reactive protein) or 3 days (whichever comes last)
- OT or IVT is no longer compatible with good clinical care (e.g. insuficient clinical effect of either OT or IVT, the patient/child is not accepting or tolerating OT or IVT.)

After this initial treatment duration, a follow-up treatment using OT with reduced dose is planned for 1 week in case of joint infection only, 4 weeks in case of spondylodiscitis and 3 weeks in case of other bone infections.

The primary clinical question of interest is: *"Is the risk of having sequelae 6 months after BJI not more than 5% higher when initiating high dose OT as compared to IVT, regardless of initiation of any additional interventions during the follow-up, including switching treatment (from OT to IVT or the other way around), when needed to ensure good clinical care?"*

Accordingly, the primary estimand is described by the following attributes:

- Population: Children and adolescents (3 months to 18 years) with suspicion of first uncomplicated BJI, possibly already treated with either OT or IVT for up to 24h and for whom the suspicion of BJI will not be abandoned within 6 months.
- Endpoint: Presence of an abnormal mobility or function of the affected joint/bone at 6 months follow-up, as evaluated by a blinded clinical examination by a qualified pediatrician and/or pediatric orthopedic surgeon.
- Treatment: The investigational interventions (initiating high-dose OT vs IVT) regardless of any subsequent treatment decision needed to ensure good clinical care ("treatment policy strategy", see [7]).
- The intercurrent event "any subsequent treatment decision" is addressed by the treatment condition of interest attribute ("treatment policy strategy", see [7]). The intercurrent event "suspicion of infection is abandoned" is addressed by the population of interest attribute ("Principal stratum strategy", see [7])
- Population-level summary: Difference in 6-month risk of sequelae between treatment conditions

Remark: the rationale for the above "population" and "intercurrent events" attributes, is the following. The population (defined above) slightly differs from that of the research question ("children with BJI" versus "children with suspicion of BJI, for whom who the suspicion of BJI will not be abandoned within 6 months"). This is because it was impossible to include children for which BJI was present with perfect certainty (due to necessity of acute inclusion) and because the population definition of the estimand was chosen to match the inclusion/exclusion criteria. However, we could not think of any reason why the population of the estimand should differ from that of the research question by more than a negligible proportion of patients, equally distributed between the two treatment arms. Hence, we think that for all practical purposes, the two populations can be considered as the same.

3.1.2 Secondary

There are two secondary outcomes:

1. Treatment failure within 28 days. It is defined by two pediatric specialists as a change of antibiotic therapy due to non-acute treatment failure. Although the two specialists are not restricted to define treatment failure, it was suggested in the protocol to define it following these parameters:
    - Temperature above 38.5°C after more than 72 hours of antibiotic therapy.
    - Increasing C-reactive protein after more than 96 hours of antibiotic therapy.
    - No improvement in mobility or pain after 120 hours of antibiotic therapy.
2. Recurrent infection within 6 months. It is defined as the occurrence of (at least one) recurrence of symptoms and signs (same anatomical location) after completion of antibiotic treatment requiring further antibiotic administration.

Correspondingly, the secondary objectives are to estimate the risks for each of the two secondary outcomes, with OT and IVT; as well as the risk differences. The clinical questions of interest are:

"What are the risks of treatment failure within 28 days when initiating high-dose OT and IVT, regardless of any additional interventions, when needed to ensure good clinical care? What is the corresponding risk difference?"

"What are the risks of recurrent infection within 6 months when initiating high-dose OT and IVT, regardless of initiation of any additional interventions during the follow-up, including switching treatment (from OT to IVT or the other way around), when needed to ensure good clinical care? What is the corresponding risk difference?"

Accordingly, the corresponding estimands are defined as the primary estimand of Sec. 3.1.1, except for the endpoint attributes, which are defined by the above items 1 and 2.

3.2 Study Design

This is a two-arm, randomized (1:1), parallel, nationwide open-label non-inferiority clinical trial. Only the assessment of the primary outcome is blinded. Randomization was stratified by C-reactive protein (< 35 vs ≥ 35 mg/L). Patients are followed up until the 12-month follow-up visit (which can take place between 11 and 14 months after randomization), although the primary outcome is defined at the 6-month follow-up visit (which can take place between 5 and 9 months after randomization). Outcomes are defined by the medical examinations at the two planned visits (6 or 12 months), clinical examinations during the course of the disease and via self-reported (electronic) questionnaires, filled in by the patients or their parents. The questionnaires provide daily and weekly patient reported outcomes.

4 Statistical Hypotheses

The null hypothesis to be tested in relation to the primary estimand (detailed in Sec. 3.1.1) is as follows:

- **Null hypothesis:** the risk of having sequelae 6 months after BJI is **at least** 5% higher when initiating high-dose OT than when initiating IVT [2]. Formally, $\mathcal{H}_0 : \pi_1 - \pi_0 \geq 5\%$, where π_1 and π_0 are the 6 months risk when initiating high-dose OT and IVT, respectively.

versus

- **Alternative hypothesis:** the risk of having sequelae 6 months after initiating BJI is **at most** 5% larger when initiating high-dose OT than when initiating IVT. Formally, $\mathcal{H}_1 : \pi_1 - \pi_0 < 5\%$.

[2] As for the definition of the estimand, it is the risk regardless of initiation of any additional interventions during the 6 months follow-up, including switching treatment (from OT to IVT or the other way around), when needed to ensure good clinical care.

4.1 Multiplicity Adjustment

No multiple testing correction will be used, as formal hypothesis testing will be performed only for the primary endpoint/estimand described in Section 3.1.1, for the null hypothesis defined in Section 4.

Reporting for other endpoints/estimands will be limited to point estimates of effects with 95% (two-sided) confidence intervals. The widths of the intervals will not be adjusted for multiplicity and therefore it will not be possible to use them in place of formal hypothesis testing. This is in line with common recommendations [10].

5 Analysis Sets

- The **"Main analysis set"** consists of all randomized participants whose suspicion of infection is not abandoned during the study period and for whom the primary endpoint is not missing. The primary endpoint is missing if the patient did not attend the follow-up after 6 months. More precisely, it is missing if the patient did not attend the "6-month follow-up visit" between 5 and 9 months after randomization.

- The **"Secondary outcomes analysis set"** consists of all randomized participants whose suspicion of infection is not abandoned during the study period. That is, the dataset obtained after adding the patients who did not attend the follow-up after 6-months to the "Main analysis set".

- The **"All participants analysis set"** consists of all randomized participants.

- The **"Long-term outcomes analysis set"** consists of all randomized participants whose suspicion of infection is not abandoned during the study period and for whom the "12 months follow-up" data are not missing. The "12 months follow-up" is missing if it did not happen between 11 and 14 months after randomization.

- The **"Principal stratum analysis set"** consists of all randomized participants whose suspicion of infection is not abandoned during the study period, for which the primary endpoint is not missing and who did not switch treatment during the initial treatment period following randomization (from OT to IVT or the other way around). That is, the set obtained after excluding from the "Main analysis set" the patients who switched during the initial treatment period.

6 Statistical Analyses

6.1 Primary Endpoint Analysis

6.1.1 Main Analytical Approach

For this analysis we will use the "Main analysis set" detailed in Section 5. We will estimate the primary estimand, that is, the difference in risk $\pi_1 - \pi_0$, as the empirical (i.e. observed) proportion of participants with sequelae after 6 months in the high-dose OT intervention group minus the same proportion in the IVT intervention group. An appropriate one-sided 97.5% confidence interval and a matching p-value for the null hypothesis $\mathcal{H}_0 : \pi_1 - \pi_0 \geq 5\%$ will be computed using an exact

unconditional test. An exact approach will be used instead of alternatives based on (asymptotic) normal approximations, because the expected number of events in each group are very low (say, 0 or 1) and thus usual (asymptotic) approximations do not seem reliable. Specifically, we will use an exact unconditional test using the score statistic ordering, as implemented in the function `uncondExact2x2` of the `R` package `exact2x2`; see [4] for the mathematical details and Appendix 8.1 for the specific `R` code that we will use. The score statistic ordering was chosen for its good power properties, as compared to alternatives [4].

If the upper limit of the 97.5% confidence interval of the risk difference $\pi_1 - \pi_0$ does not include the non-inferiority margin 5% or, equivalently, if the matching p-value is less than 2.5%, we will reject the null hypothesis[3].

In addition, we will report the estimates of the risk in each group (i.e. the empirical proportions) together with exact binomial two-sided 95% confidence intervals (computed using the `binom.test()` function of `R`).

Note that the above statistical analyses will ignore that the randomization was stratified by C-reactive protein (< 35 vs ≥ 35). This is usually not recommended, as this usually leads to a decrease in power [5, 8]. However, our context is very specific, as we expect close to zero sequelae in each group (whatever the C-reactive protein level). This has two important consequences. First, the loss of power is expected to be non-existent or negligible. Second, ensuring proper type-I error control using a "covariate-adjusted" alternative approach seems difficult using standard software.

Another important remark also needs to be made, about the rationale for how we deal with the intercurrent event "the suspicion of BJI is abandoned". Although it is generally agreed that "*the subset of subjects who experience an intercurrent event [here: suspicion of BJI is abandoned] on the test treatment will often be a different subset from those who experience the same intercurrent event on control*"; and therefore that *"Treatment effects defined by comparing outcomes in these subsets confound the effects of the different treatments with the differences in outcomes possibly due to the differing characteristics of the subjects."* [7] (see also [2] for similar discussions); we do not believe that this is the case in our study (as already mentioned above, in the remark at the end of Section 3.1.1). Therefore, we do not suspect the planned analysis to be biased. Indeed, we think that for each patient for whom the suspicion of BJI is not abandoned, it would not have been abandoned either if the child had been randomized to the other treatment arm. This seems reasonable as the main reason to abandon the suspicion (imaging, blood samples or clinical follow-up reveals another diagnosis) is not expected to be a consequence of the received intervention (OT or IVT). However, in the rare case where the diagnosis remains uncertain, there is a small risk that the treating clinician could abandon the diagnosis of BJI (and thereby ending the antibiotic treatment) in the IVT group more often than in the OT group. This is because IVT is more invasive than OT and therefore the incentive to stop IVT could be stronger than the incentive to stop OT, in case of doubt about the BJI diagnosis. We believe this risk to be negligible and it has been accepted in the overall design of the study, as a minor limitation.

6.1.2 Sensitivity Analyses

The main analysis is based on the "Main analysis set", which do not include patients for whom the primary endpoint is missing. This choice is thought to be conservative as it is considered very unlikely that a patient with sequelae at 6-months would not attend the follow-up visit at 6-months.

[3] Note that the `uncondExact2x2` function guaranties compatible inference via either confidence interval or p-values, so concluding based on the p-value or the confidence interval is fully equivalent here (unlike when using some other software/methods).

This is especially unlikely as telephone interviews focusing on signs of sequelae were planned for all patients not attending the 6 month follow-up.

Nevertheless, we will perform the following sensitivity analysis, for completeness. We will report the results (p-value and upper limit of the 97.5% CI) for all possible combinations of the values of missing outcomes in the OT and IVT groups. An graphical representation similar to the "Enhanced tipping-point display" presented in Liublinska et al. [9, Fig. 2] will be reported, to provide an overview of the results. It will show a heat map of the upper limit of the 97.5% CI (or p-value) for each possible combinations.

As it is considered very unlikely that a patient with sequelae at 6-months would not attend the follow-up visit at 6-months, the results of the sensitivity analysis assuming that none of the patients with missing data had sequelae will be of special interest.

6.1.3 Supplementary analysis: principal stratum estimand

The following clinical question is also of interest: *"Is the risk of having sequelae 6 months after BJI not more than 5% larger when initiating high-dose OT as compared to IVT, regardless of initiation of any additional interventions during the follow-up, in the principal stratum of patients who will never need to switch treatment (from OT to IVT or the other way around) to ensure good clinical care?"*

The corresponding "principal stratum" estimand is described by the following attributes:

- Population: subgroup of population of the primary estimand (See Section 3.1.1), which includes only patients who will never need to switch treatment (from OT to IVT or the other way around) to ensure good clinical care.
- Endpoint: same as for the primary estimand (Sec. 3.1.1)
- Treatment: same as for the primary estimand (Sec. 3.1.1)
- The intercurrent event "treatment switch (from OT to IVT or the other way around)" is addressed by the population of interest attribute ("Principal stratum strategy", see [7]) and "any other subsequent treatment decision needed to ensure good clinical" is addressed by the treatment condition of interest attribute ("treatment policy strategy", see [7]).
- Population-level summary: same as for the primary estimand (Sec. 3.1.1)

Rationale for considering this estimand: First, it was the analysis that the main investigators had in mind when referring to the "per-protocol" analysis in the protocol paper [11]. As this analysis was pre-specified in the protocol, it should be performed, hence it is included in this SAP. Second, this is a non-inferiority study and of course a lot of treatment switching could be suspected to make the results of the two groups more similar, increasing the chance of showing non-inferiority. However, fundamentally the research hypothesis of this project is that high-dose OT is non inferior to IVT in the way the antibiotics are processed by the patient, hence giving similar clinical outcomes. Third, it could be hypothesized that some subgroups of patients (e.g., teenager vs young children) are more or less prone to treatment switching (e.g., from OT to IVT due to vomiting). Therefore, the results of the principal stratum analysis could provide reasonable estimates of the results that could be expected in such subgroups (to some extent).

For the corresponding statistical analysis, we will proceed as for the main analysis detailed in Section 6.1.1, except that the "Principal stratum analysis set" will be used instead of the "Main analysis set".

The rationale for this statistical analysis is the following. Although it is generally agreed that *"the subset of subjects who experience an intercurrent event [treatment switch in our study] on the test treatment will often be a different subset from those who experience the same intercurrent event on control"*; and therefore that *"Treatment effects defined by comparing outcomes in these subsets confound the effects of the different treatments with the differences in outcomes possibly due to the differing characteristics of the subjects."* [7], we do not believe that this is the case in our study. Therefore, we do not suspect the planned analysis to be biased. The rationale behind this is the following.

Let Y and S denote the outcome and indicator of treatment switch, respectively, and Y^x and S^x the corresponding counterfactuals for each intervention $x = 0$ (IVT) and 1 (OT); we expect

$$\underbrace{E(Y^1 - Y^0 | S^0 = S^1 = 0)}_{\text{principal stratum estimand}} = \underbrace{E(Y^1 | S^1 = 0) - E(Y^0 | S^0 = 0)}_{\text{estimand of subset analysis}}$$

simply because

$$E(Y^1 | S^0 = S^1 = 0) = E(Y^1 | S^1 = 0) \tag{1}$$

$$\text{and} \quad E(Y^0 | S^0 = S^1 = 0) = E(Y^0 | S^0 = 0)\ . \tag{2}$$

Assumption (1) above is fulfilled if the risk of sequelae is the same among patients randomized to OT who will never switch to IVT, regardless of whether they would have switched to OT if they had been randomized to IVT. Similarly, Assumption (2) is fulfilled if the risk of sequelae is the same among patients randomized to IVT who will never switch to OT, regardless of whether they would have switched to IVT if they had been randomized to OT. We believe these assumptions reasonable as the main reasons for switching (e.g, vomiting that prevents OT) do not seem to share a common cause with the outcome of interest (sequelae); or, at least, not a common cause which has a non-negligible effect on the outcome.

6.2 Secondary Endpoints Analyses

The main statistical analysis for the secondary outcomes and estimands defined in Section 3.1.2 will be similar to that of the primary outcome defined in Section 6.1.1, except that:

- we will compute a two-sided 95%-CI instead of a one-sided 97.5%-CI and we will not compute an accompanying p-value (see appendix 8.3 for the specific `R` code that we will use).
- we will use the "Secondary outcomes analysis set" instead of the "Main analysis set" (see details of the sets in Section 5).

The change of analysis set is because we want to use as much data as possible for the analyses and because we cannot think of any reason why a missing follow-up at 6 months (i.e., a missing primary outcome) should lead to excluding a patient for the secondary outcome analyses. For the same reason, we will proceed similarly for the tertiary (safety and exploratory) outomes (see section 6.3).

6.3 Tertiary (safety and exploratory) Endpoints Analysis

We will also analyze the following Safety **(S)** and Exploratory **(E)** outcomes listed below. The information in parenthesis at the end of the description of each outcome refers to the statistical methods that will be used to analyze them, which is described below. Note that S1–S3 and E1–E5 were already listed in Table 1 in the protocol paper [11].

S1: severe complications during antibiotic treatment within 28 days. E.g., need for intensive care, septic shock, organ failure, pyomyositis, endocarditis, deep venous thrombosis. **(B + TTE)**

S2: need for surgical intervention during antibiotic treatment within 28 days; excluding diagnostic surgical intervention (diagnostic joint aspiration or diagnostic bone biopsy). **(B + TTE)**

S3: non-severe treatment-related adverse events within 3 months. E.g., complications of intravenous access (infection, need for replacement, extravasation) and drug side effects reported by medical staff, patients or parents (electronic questionnaire). **(B + TTE)**

E1: time from initiation of antibiotic treatment to apyrexia within 28 days, as reported by the patients or parents. Apyrexia will be defined by a temperature $< 38\,^{\circ}\text{C}$ for 24 hours. Patients or parents were asked to report the temperature three times per day (until apyrexia). **(KM)**

E2: total duration of antibiotic therapy within 3 months. **(Q + TTE)**

E3: sequelae within 12 months. E.g., abnormal mobility and growth abnormalities, assessed by clinical examination by a qualified pediatrician 12 months after the initiation of treatment, accepted range 11–14 months. **(B)**

E4: radiological abnormalities within 12 months. It is assessed by a qualified radiologist 12 months after initiation of treatment, accepted range 11–14 months. **(B)**

E5: secondary infection with antimicrobial-resistant organisms or Clostridioides difficile, within 3 months. **(B)**

E6: discontinuation of the randomized treatment strategy within 28 days. **(B + TTE)**

E7: early termination of falsely included patient within 6 months (i.e., suspicion of infection is abandoned). **(B + TTE)**

E8: time to shift to 'follow-up treatment' after clinical improvement (pain and mobility) and paraclinical improvement (decreasing C-reactive protein) within 3 months. **(Q + TTE)**

E9: duration of 'follow-up treatment' within 3 months (Note: E8+E9= E2). **(Q + TTE)**

We prespecify the analysis sets and the statistical methods that will be used to analyze each of the above Safety and Exploratory outcomes as follows. Additional (post hoc) exploratory analyses might be performed.

Analysis sets. For the analyses of outcomes S1, S2, S3 and E7 we will use the "All participants analysis set", whereas we will use the "Secondary outcomes analysis set" for outcomes E1-E2 and E5-E6 and E8-E9 and the "Long-term outcomes analysis set" for outcomes E3-E4. This information is summarized in Table 6.3.

	Analysis set				
Outcome	"Main"	"Second"	"All"	"Late"	"PS"
Primary	X				X
S1			X		
S2			X		
S3			X		
E1		X			
E2		X			
E3				X	
E4				X	
E5		X			
E6		X			
E7			X		
E8		X			
E9		X			

Note: "Main", "Second", "All", "Late" & "PS" stand for "Main analysis set", "Secondary outcomes analysis set", "All participants analysis set", "Long-term outcomes analysis set" and "Principal stratum analysis set".

Table 1: Analysis sets for each outcome.

Binary outcomes analyses (B). Outcomes S1–S3, E3-E7 are binary outcomes. They will be analyzed using the same statistical method as for the primary outcome, except that we will compute a two-sided 95% CI instead of a one-sided 97.5% and that no p-value will be computed (see appendix 8.3 for the specific `R` code that we will use).

Time-to-event analyses (TTE). Outcomes S1-S3, E1–E2 and E6-E9 are –or can also be analyzed as– time to event outcomes. They will be analyzed by estimating (and plotting) the cumulative incidence/risk function within each treatment arm, with pointwise two-sided 95%-CI. That is, we will present the proportion of event observed before day t, for $t = 0, 1, 3, \ldots$ for each treatment arm and the corresponding exact binomial two-sided 95% CIs (computed using the `binom.test()` function of `R`). This simple analysis is planned because we expect no censoring.

Quantitative outcomes analyses via quantiles (Q). Outcomes E2, E8 and E9 are quantitative outcomes (duration). They will (also) be analyzed by estimating the median within each treatment arm and their difference, together with two-sided 95%-CIs. The CI will be computed using non-parametric BCa Bootstrap (using B=10,000 replicates, via the `R` packages `boot` and `simpleboot`, see appendix 8.2 for the specific `R` code that we will use).

Kaplan-Meier analysis (KM). Outcome E1 is a time-to-event outcome (time to apyrexia). It will not be fully observed for all patients due to missing data, which will lead to censored observations. Hence we will use the Kaplan-Meier method to compute the "survival curves" for the event "apyrexia" within each group. We will define the time to apyrexia as the time to the third consecutive recording (observed or imputed) with a temperature < 38 °C (i.e., approximately 24h with a temperature below 38 °C). Missing temperature recordings will lead to imputation or censoring as follows. Last observation carried forward will be applied to impute up to two consecutive missing temperature recordings. If more than two consecutive recordings are missing, we will censor the

time to apyrexia at the time of the third consecutive missing temperature (the first time three consecutive temperatures are missing, if this happens several times). The `prodlim` package of `R` will be used to compute the Kaplan-Meier curves in each arm with pointwise two-sided 95%-CI. The corresponding median survival time in each arm and their difference will also be computed, also with two-sided 95%-CI, computed using non-parametric BCa Bootstrap (using B=10,000 replicates, via the `R` packages `boot` and `prodlim`, see appendix 8.4 for the specific `R` code that we will use).

Further remarks: for the outcome E6 (discontinuation of the randomized strategy), data about the reason for discontinuation have been collected. Additional (post hoc) exploratory analyses of these additional data will be performed too.

6.4 Other analyses

6.4.1 CRP analysis

Data about C-reactive protein (CRP) level measured via blood samples during the follow-up will also be reported. Descriptive statistics and graphical representation will summarize the data observed within each treatment arm at each follow-up day after randomization. Boxplots (per arm and per day), dotplots and / or spaghetti plots will be considered and produced as relevant. Descriptive statistics about the timing and number of CRP measurements per patients will also be summarized within each arm. Further (post hoc) exploratory analyses of the CRP data will be considered.

6.4.2 Level of mobility and pain

Data about level of mobility and pain within 14 days of follow-up will also be reported. The data come from daily grading of symptoms by medical staff and daily standardized pain scores from participants and/or parents. These two score systems were used: Visual Analogue Scale and Face Legs Activity Cry Consolability scale. Both scores range from 0 (no pain) to 10 (worst pain). As for CRP data, descriptive statistics and graphical representation will summarize the data observed within each treatment arm, during the follow-up. Boxplots (per arm and per day), dotplots and/or spaghetti plots will be considered and produced as relevant. Details about the number of missing values will also be summarized within each arm. No specific statistical inference method is pre-specified to analyze these data, as it was thought difficult to pre-specify a meaningful analysis before performing exploratory analyses of the missing data (to better understand their complexity and whether they might lead to substantial biases). Post hoc analyses will be considered and any "finding" from such exploratory analyses will be reported as such (i.e., not coming from a pre-specified analysis).

6.4.3 Subgroup analyses

We plan to perform subgroup analyses for these subgroups:

1. patients who did not initiate antibiotics treatment (for up to 24h) before inclusion.[4]
2. patients with confirmed infection with staphylococcus aureus (positive culture or PCR from blood, joint fluid or bone biopsy).

[4] **Reminder:** inclusion criteria allowed patients to be already treated with either OT or IVT for up to 24h before inclusion/randomization. See Section *"Delayed inclusion"* in the protocol paper [11].

3. patients with confirmed infection with any bacteria (positive culture or PCR from blood, joint fluid or bone biopsy).

4. patients with spondylodiscitis.

For each subgroup, the main analysis of Section 6.1.1 will be replicated as well as that of the secondary outcomes of Section 6.2. Limited power is expected in some subgroups, but the number of events observed in each group will be interesting and informative to report anyway.

6.4.4 Adherence

Data about number of doses missed per week have been collected and will be reported, via simple descriptive statistics, for each arm and each of the two treatment periods: "initial treatment" (randomized to OT or IV) and "follow-up treatment" (oral) after clinical and paraclinical improvement. Note that important data about adherence will also be reported via the analysis of tertiary exploratory outcomes E6, defined in Section 6.3.

6.4.5 Timing of follow-up visits

Descriptive statistics, per arm, will be presented for the time from randomization to the "6-month" and "12-month" follow-up visits (which could not occur after exactly 6 and 12 months for all patients). Median, minimum, maximum, first and third quartiles will be reported.

6.4.6 Recruitment

Recruitment of the patients will be summarized via descriptive statistics. Especially, start and end dates of recruitment will be presented as well as a flowchart, inspired by the CONSORT guidelines and template [13].

6.4.7 Descriptive statistics of baseline characteristics

Baseline characteristics will be descriptively summarized, within each of these four groups: randomized to OT and to IVT within the main analysis set and to OT and to IVT among early terminated patients (i.e., among those for whom suspicion of infection is abandoned). The list of baseline variables to be summarized includes:

- Age
- Sex
- Anatomical location of the infection
- Time from first symptom to start of treatment
- Clinical presentation (e.g., fever, pain, local signs of infection)
- CRP
- Imaging

- Microbiological investigations

For quantitative variables, we will present median, first and third quartiles and also minimum and maximum. Categorical variables will be summarized by numbers and percentages. Hypothesis tests will not be performed to compare baseline characteristics, but clinical importance of any imbalance will be noted. This is in line with usual recommendations [12].

6.4.8 Descriptive statistics of outcomes

Outcomes (primary, secondary and tertiary) will also be descriptively summarized, within relevant subgroups. Similar descriptive statistics as for baseline characteristics will be used. Mean and standard deviation will be presented instead of median, first and third quartiles (or in addition to those) whenever appropriate. For each outcome, the relevant subgroups include the OT and IVT subgroups of the analysis set used for the main statistical analysis of this outcome (detailed in Table 6.3).

6.4.9 Screening data

Screening data, about assessment for eligibility, have been collected. They will be presented in a in a flow diagram, following the CONSORT guidelines [13].

6.5 Changes to protocol-planned Analyses

- In this document, we clearly state that the 6-month follow-up visit is accepted to occur at any time between 5 and 9 months after randomization. This detail was missing in the protocol paper [11], although a similar information was present for the 12-month follow-up visit (accepted range 11-14 months).

- In the Section "Missing data" of the protocol paper [11], about the main analysis of the primary outcome, it was written: *"If the missing data are more than 5%, we will impute data based on available knowledge for the patient and the observations of the other patients in the same randomized group."*. Further thinking made us change to a "complete case" data analysis detailed in Section 6.1.1. It is thought as a more conservative and trustworthy analysis and it will be complemented by the sensitivity analysis detailed in Section 6.1.2.

- A per-protocol analysis was mentioned in the protocol paper [11]. In this document, this analysis is not referred to as the "per-protocol analysis" but as the supplementary analysis corresponding to the "principal stratum estimand", in Section 6.1.2. This is our attempt to closely follow the recent guidelines about estimands [7].

- The term "principal stratum" refers to a different subgroup in this document (see Section 6.1.3) and in the protocol paper [11]. The version of this document was thought as more natural after we detailed the main estimand in Section 3.1.1.

- The tertiary exploratory outcomes E6-E9 were not explicitly mentioned in the protocol (except maybe for E8 in the IVT arm, see next item).

- The exploratory outcome E8 in the IVT arm is the duration of intravenous antibiotics. Page 7 of the protocol paper [11] it is written that it will be reported but that it is not considered as an "outcome". In this SAP, we define it as a tertiary exploratory outcome. This is only for convenience, as it will be analyzed as other tertiary exploratory outcomes.

- Related to the analysis method for outcome S3, in section *"Statistical analyses for the safety outcomes"* of the protocol paper [11], it is written that the number of *"days the child is affected by one or more of these non-serious adverse events"* will also be reported for each arm and compared between arms. Unfortunately, the necessary data to perform this analysis have not been collected. This is because data were collected from the patients or their parents for each adverse event separately. Hence, based on the available data we cannot identify when two adverse events occurred within the same day or at two different days.

- Level of mobility and pain was considered as an "exploratory" outcome in the protocol paper [11], whereas in this SAP we present the corresponding analysis plan in the "Other analyses" section (in subsection 6.4.2). This is because no specific statistical inference method is pre-specified to analyze the data about level of mobility and pain (unlike for the other exploratory outcomes).

7 Sample size determination and power calculation

As stated in Nielsen et al. [11], the sample size of 180 children with confirmed BJI (90 in each group) was computed to provide 90% power assuming: a non-inferiority margin of 5%, a risk of sequelae of 1% in both group and 10% dropout rate. The usual asymptotic normal approximations was used, i.e.,

$$\text{Power} \approx \Phi\left(\frac{(0.99-0.99)+0.05}{\sqrt{0.99\times 0.01/81+0.99\times 0.01/81}}-1.96\right)\approx 90\%$$

where $n = 90 \times (1 - 0.1) = 81$ subjects are expected fully observed (with no dropout) in each group and Φ is the cumulative standard normal distribution function, see e.g. Chow et al. [1, Sec. 4.2.2].

Because the expected number of events in each group are very low (0 or 1) and because we will use an exact test (see Sec 6.1 above), the initial (above) asymptotic power calculation has been suspected to not be sufficiently precise. Hence, an alternative exact power computation was performed and showed that with $n = 81$ subjects in each group, the power is actually approximately 51%, 68% or 82% if we expect 1%, 0.5% or 0.25% risk of sequelae in both group[5]. This suggests that the study is substantially underpowered if there is indeed 1% risk of sequelae in both group, but decently powered if the risk is slightly lower (equal to 0.25%), which is not unrealistic. These additional results complement those presented in the protocol paper [11]. They are the consequence of additional, more recent, thinking.

[5] The computation was performed by computing the likelihood of observing $(x_1, x_0) = (0,0), (0,1), (1,0), (1,1), \ldots$ and summing-up the likelihood of each case leading to a significant result, where (x_1, x_0) denote the number of sequelae in the OT and IVT groups.

8 Supporting Documentation

8.1 R code for the exact unconditional test (Primary Estimand Analysis)

```
library(exact2x2)
uncondExact2x2(x1=x1obs,              # x1obs is the number of sequelae observed in IVT group
               n1=n1obs,              # n1obs is the number of patients in IVT group
               x2=x2obs,              # x2obs is the number of sequelae observed in OT group
               n2=n2obs,              # n2obs is the number of patients in OT group
               parmtype="difference", # estimand is risk difference
               alternative = ("less"),# one-dided test (non-inferiority)
               method="score",        # score statistic ordering is used
               conf.level = 0.975,    # 97.5% confidence level
               nullparm=0.05,         # non-inferiority margin is 5%
               conf.int=TRUE)         # computation of confidence interval wanted
```

8.2 R code: 95%-CI for the median and median difference via BCa bootstrapping

```
library(boot)
library(simpleboot)
#--
x0  # vector of observed outcomes in group 0
x1  # vector of observed outcomes in group 1
set.seed(20231110) # set the seed
# Median difference 95% CI
resDiff <- two.boot(x0, x1, median, R = 10000)
boot.ci(resDiff,conf = 0.95,type="bca")
# Median in group 0
res0 <- one.boot( x0, median, R = 10000)
boot.ci(res0,conf = 0.95,type="bca")
# Median in group 1
res1 <- one.boot( x1, median, R = 10000)
boot.ci(res1,conf = 0.95,type="bca")
```

8.3 R code for exact 95%-CI for the difference in risk

```
library(exact2x2)
uncondExact2x2(x1=x1obs,                  # x1obs is the number of events in IVT group
               n1=n1obs,                  # n1obs is the number of patients in IVT group
               x2=x2obs,                  # x2obs is the number of events in OT group
               n2=n2obs,                  # n2obs is the number of patients in OT group
               parmtype="difference",     # estimand is risk difference
               alternative = "two.sided", # two-sided CI wanted
               method="score",            # score statistic ordering is used
               conf.level = 0.95,         # 95% confidence level
               conf.int=TRUE)             # computation of CI wanted
```

8.4 R code for 95%-CI for median survival time per arm and their difference

```
rm(list=ls())
library(boot)
library(prodlim)
# Prepare Data
```

```
d <- data.frame(time=Times,status=Status,arm=Arm)
# Times: times from initiation of antibiotic treatment to apyrexia or censoring
# Status: status indicators (0= censored, 1=apyrexia)
# Arm: indiator of treatment arm (1=IVT, 0=OT)
#
# Helper function to Bootstrap
mediansurvDiff  <-  function(dataFrame, indexVector) {
    # resample per arm
    d1 <- subset(dataFrame[indexVector, ], dataFrame[indexVector, "arm"] == 1)
    d0 <- subset(dataFrame[indexVector, ], dataFrame[indexVector, "arm"] == 0)
    # Kaplan-Meier per arm
    KMfit0 <- prodlim(Hist(time,status)~1,data=d0)
    KMfit1 <- prodlim(Hist(time,status)~1,data=d1)
    # resulting quantiles and difference
    m0 <- quantile(KMfit0,q=0.5)$quantile
    m1 <- quantile(KMfit1,q=0.5)$quantile
    m  <-  m1 - m0
    return(c(m,m1,m0))
}
# set the seed
set.seed(20231110)
# Bootstrap resampling
totalBoot  <-  boot(d, mediansurvDiff, R = 10000, strata = d[,"arm"])
# Median time difference 95% CI
boot.ci(totalBoot,conf = 0.95,type="bca",index=1)
# Median time 95% CI in arm=1
boot.ci(totalBoot,conf = 0.95,type="bca",index=2)
# Median time 95% CI in arm=0
boot.ci(totalBoot,conf = 0.95,type="bca",index=3)
```

References

[1] Chow, S.-C., Shao, J., Wang, H., and Lokhnygina, Y. (2008). *Sample size calculations in clinical research, Second Edition*. CRC press.

[2] Colantuoni, E., Scharfstein, D. O., Wang, C., Hashem, M. D., Leroux, A., Needham, D. M., and Girard, T. D. (2018). Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*, 360.

[3] Evans, S. and Ting, N. (2015). *Fundamental concepts for new clinical trialists*. CRC Press.

[4] Fay, M. P. and Hunsberger, S. A. (2021). Practical valid inferences for the two-sample binomial problem. *Statistics Surveys*, 15:72–110.

[5] FDA (2023). Adjusting for covariates in randomized clinical trials for drugs and biological products guidance for industry. Technical report, Food and Drug Administration, https://www.fda.gov/media/148910/download.

[6] Gamble, C., Krishan, A., Stocken, D., Lewis, S., Juszczak, E., Doré, C., Williamson, P. R., Altman, D. G., Montgomery, A., Lim, P., et al. (2017). Guidelines for the content of statistical analysis plans in clinical trials. *Jama*, 318(23):2337–2343.

[7] ICH E9 (R1) (2017). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Technical report, EMA/CHMP/ICH, `www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline`.

[8] Kahan, B. C. and Morris, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*, 31(4):328–340.

[9] Liublinska, V. and Rubin, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in Medicine*, 33(24):4170–4185.

[10] NEJM (2023). Statistical Reporting Guidelines of the New England Journal of Medicine, section Multiplicity considerations. https://www.nejm.org/author-center/new-manuscripts. Accessed: 2023-10-24.

[11] Nielsen, A. B., Borch, L., Damkjaer, M., Glenthøj, J. P., Hartling, U., Hoffmann, T. U., Holm, M., Rasmussen, A. H., Schmidt, L. S., Schmiegelow, K., et al. (2023). Oral-only antibiotics for bone and joint infections in children: study protocol for a nationwide randomised open-label non-inferiority trial. *BMJ open*, 13(6):e072622.

[12] Schulz, Altman, and Moher, for the CONSORT Group (2023). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. https://www.goodreports.org/reporting-checklists/consort/. Accessed: 2023-11-17.

[13] Schulz, K. F., Altman, D. G., and Moher, D. (2010). Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and pharmacotherapeutics*, 1(2):100–107.

[14] Stevens, G., Dolley, S., Mogg, R., and Connor, J. T. (2023). A template for the authoring of statistical analysis plans. *Contemporary Clinical Trials Communications*, page 101100.