Statistical Analysis Plan (SAP)

SAP version number:	2025-01-17
Person writing the SAP:	Paul Blanche
Principal investigators:	Emma Malchau Carlsen & Ulrikka Nygaard
Protocol title:	Individualized Duration of Antibiotic Treatment in Early Onset Infection in Newborns (DurATi-n)
ClinicalTrials.gov ID:	NCT05329701

Note: inspiration to write this SAP came from the SAP template provided by TransCelerate [25] as well as recommendations from Gamble et al. [12], Stevens et al. [23] and Evans and Ting [8].

Contents

1 Statistical analysis plan approval signature page

Person writing the SAP:

Name: Paul Blanche

Position/Title: Associate Professor in Biostatistics

[redacted] February 6, 2025

Signature & Date

Approved by:

Name: Emma Malchau Carlsen

Position/Title: Senior Registrar, Associate Professor, MD, PhD

[redacted] February 8, 2025

Signature & Date

Name: Ulrikka Nygaard

Position/Title: Senior Consultant, Associate Professor, MD, PhD

[redacted] February 6, 2025

Signature & Date

2 List of Abbreviations

- CI: Confidence Interval
- CRP: C-reactive protein
- GBS: Group B streptococcus
- ITT: Intention-To-Treat
- PP: Per Protocol
- SAE: Serious Adverse Event
- SAP: Statistical Analysis Plan

3 Introduction

This document is the statistical analysis plan (SAP) for an open label, investigator-initiated, parallel group, 1:1 randomized, multicenter, non-inferiority clinical trial of "individualized" versus "standard" duration of antibiotic treatment for culture negative early-onset infection in term newborns. The individualized treatment duration is based on structured clinical assessment of symptoms and level of C-reactive protein (CRP). The individualized treatment duration can be as short as two days while the standard treatment duration is seven days. It is hypothesized that an individualized treatment duration will shorten the duration of antibiotic therapy in newborns with early onset infection with no or very limited increase in the risk of relapse.

The trial recruited newborns from all Departments of Neonatology in Denmark except one (17 out of 18). Stratified randomization was planned 36 to 48 hours after the treatment started using permuted blocks of size six, stratified on maximum CRP level within the last 48h above or below 90 mg/l. Patient accrual started in Spring 2022 and was anticipated to stop in January 2025, after inclusion of 488 newborns.

The changes to the protocol-planned analyses are summarized in Section 12. We do not consider them as major changes and the rationale for each of them is provided.

4 Objectives, Endpoints, and Estimands

4.1 Primary objectives, endpoints and estimands

There are two co-primary objectives and therefore two co-primary outcomes in this trial.

The first objective is to show that an "individualized" antibiotic treatment duration is non-inferior to a "standard" treatment duration for treating culture negative early-onset infection in term newborns, with respect to the risk of readmission due to bacterial infection. By non-inferior, we mean that the risk of readmission due to bacterial infection within 28 days is not larger with the "individualized" duration than with the "standard" duration, by more than the pre-specified non-inferiority margin of 4%. The rationale for choosing this margin is provided in Section 4.1.2.

The second objective is to show that the "individualized" antibiotic treatment duration leads to less antibiotics days than the "standard" treatment duration. More specifically, the second objective is to show superiority of the "individualized" treatment duration with respect to the

median of the total duration of antibiotics use within 28 days from treatment start. Although the initial treatment duration will systematically be shorter using the "individualized" duration, relapses can happen, leading to new initiations of antibiotic treatment. Hence it remains to confirm that the median of the total duration of treatment within 28 days is shorter using the "individualized" initial duration than using the "standard" initial duration. This is precisely the second objective.

The two treatment strategies (interventions) being investigated for initial treatment are detailed below. Of note, the two strategies below are defined for newborns who have already received 36 to 48 hours of treatment, hence none can lead to a treatment discontinuation before a newborn has received at least 36 hours of antibiotic treatment.

- **Experimental ("individualized")**: discontinuation of antibiotics when both of the following two criteria are fulfilled:

  1. The infant has been 24 hours without clinical symptoms of infection, after systematic clinical evaluation by a neonatologist. The clinical symptoms that are considered are those specified in Table 1 below. They are those of the "easy-to-use" scoring system originally published by Stocker et al. [24], up to minor modifications.
  2. CRP is declining and $\leq$ 30 mg/l.

  We define CRP as declining if the current value is less than the most recent previous value. CRP is assessed regularly[1] until the condition of the previous item is fulfilled and once every 24-48 hours afterwards.

- **Control ("standard")**: discontinuation of antibiotics after the same standard duration for all infants. This standard duration is set to seven days in nearly all participating Departments (and in all of the largest), but it is five days in a few[2]. In the extremely unlikely case where a child would still have clinical symptoms of infection by the end of the standard duration, treatment will be prolonged as needed. Hence, using a "standard" treatment duration strategy cannot result in a shorter initial treatment duration than using the "individualized" treatment duration strategy.

1	Tachypnea, apnoea, use of respiratory support or oxygen supplementation
2	Tachycardia
3	Arterial hypotension
4	Hypothermia or hyperthermia
5	Seizure, floppy infant, irritability, or lethargy
6	Feeding insufficiency, vomiting, or ileus

Table 1: Clinical symptoms (slightly modified from Stocker et al. [24, Fig. 1]).

4.1.1 Readmission due to bacterial infection

The first co-primary outcome (**P1**) is hospital readmission due to bacterial infection (or death, although none is expected), within 28 days after treatment started. Specifically, hospital readmission due to bacterial infection is defined as hospital readmission that fulfills all of the following:

[1] By regularly we mean every 18 to 48h, typically every 24h, approximately.

[2] Specifically, seven days in 15 out of 17 departments, including all of the largest; five days in two departments.

- ≥ 1 clinical symptom(s), among those listed in Table 1
- CRP during readmission > 10 mg/l
- ≥ 72 hours of antibiotic treatment received following readmission

Note that there is no minimum time for the time to readmission as we define and record it. A newborn can be readmitted as early as the day of discharge or the following day.

The first co-primary clinical question of interest is: *"Is the risk of hospital readmission due to bacterial infection within 28 days after the start of treatment not more than 4% higher when using the individualized treatment duration than when using the standard duration of seven days, regardless of initiation of any additional interventions during the follow-up, when needed to ensure good clinical care?"*

To answer this question, we will use two co-primary estimands. The first corresponds to an **"intention-to-treat estimand"** and is described by the following attributes:

- Population: infants diagnosed with a probable or possible infection within 0-72 hours from birth according to a structured infection risk assessment, who have received 36 to 48h of antibiotic treatment and no positive blood culture result so far; infants are born with gestational age $\geq$ 35 weeks and weight $\geq$ 2000 g.
- Endpoint: hospital readmission due to bacterial infection (as defined above; or death) within the first 28 days after treatment initiation.
- Treatment: The investigational interventions ("individualized" vs "standard", as defined above) regardless of any subsequent treatment decision or intervention during the follow-up needed to ensure good clinical care and regardless of adherence ("treatment policy strategy", see [15]).
- The intercurrent event "any subsequent treatment decision or intervention" is addressed by the treatment condition of interest attribute ("treatment policy strategy", see [15]). The intercurrent event "death within 28 days" is addressed by the endpoint definition ("composite variable strategy", see [15]). The intercurrent event adherence to the investigational intervention (i.e., "the patient actually received the treatment according to the randomized treatment strategy", i.e., neither shorter nor longer duration than randomized duration) is also addressed by the treatment condition of interest attribute ("treatment policy strategy", see [15]). There are no other relevant intercurrent events anticipated.
- Population-level summary: Difference in risk between treatment conditions.

The second co-primary estimand corresponds to a **"per protocol estimand"** and only differs from the previous estimand for these attributes:

- Treatment: The investigational interventions ("individualized" vs "standard", as defined above) regardless of any subsequent treatment decision or intervention during the follow-up needed to ensure good clinical care ("treatment policy strategy", see [15]). Unlike with the previous estimand, here adherence to the randomized treatment strategy is required ("hypothetical strategy", see [15]).

- The intercurrent event "any subsequent treatment decision or intervention" is addressed by the treatment condition of interest attribute ("treatment policy strategy", see [15]). The intercurrent event "death within 28 days" is addressed by the endpoint definition ("composite variable strategy", see [15]). Another possible intercurrent is "the patient did not receive the randomized treatment strategy" (e.g., stopped treatment prematurely or continued too long). A scenario is envisaged in which this does not occur ("hypothetical strategy", see [15]). There are no other relevant intercurrent events anticipated.

Rationale for these co-primary estimands. As we expect a very small proportion of non adherence ($< 5\%$), the two estimands are considered very similar. However, in general, it is often considered *"advantageous to demonstrate a lack of sensitivity of the principal trial results to alternative choices of the set of subjects analysed"* [14, Sec 5.2.3]. An important advantage of the "Intention-To-Treat" estimand over the "Per Protocol" estimand is that it can be estimated without bias under weaker assumptions, because of randomization [11]. However, the "Intention-To-Treat" estimand interpretation depends on the level of adherence and poor adherence will generally make the two arms more similar, hence facilitating demonstration of non-inferiority. Specifically, in our trial, a non negligible proportion of newborns randomized to "short" receiving longer treatment duration than according to the "short" treatment strategy could make the two arms more similar than they should, according to the randomization. The interpretation of the "per-protocol" estimand does not depend on the level of adherence, but inference for this estimand inevitably relies a few additional assumptions (e.g. unmeasured confounding), which might be questionable [13, 2]. Consequently some guidelines recommend that both should have equal importance [5]. See e.g., Evans and Ting [8, Sec. 8.3.2] for similar remarks.

4.1.2 Non inferiority margin choice

The rate of relapse after a proven GBS infection has not been extensively studied, but at time of study planning it was estimated to be between 1% and 8%. Further, two epidemiological studies reported recurrent infection in 1% of cases with proven early onset infection [26, 7]. Only very few studies have examined the rate of relapse in babies with culture-negative infection before the present study was initiated, but it was estimated to be between 0.5% and 1%. The non-inferiority margin was set to 4%, thus accepting 4% more recurrence of infection in the intervention group who received an individualized duration of antibiotic treatment. The non-inferiority margin is justified by the follow-up of babies receiving the individualized treatment, with free access to the neonatal department, thus enabling early recognition of a potential relapse and balanced against the benefit associated with a shorter duration of treatment. A shorter duration of treatment reduces the antibiotic-associated gastrointestinal side effects and will, depending on how the antibiotics are administrated, in some cases, also reduce the length of hospital stay. Additionally, if the non-inferiority margin were set lower, it would require a larger sample size, increasing the cost of the study and its duration. This would have challenged the feasibility of the study.

4.1.3 Total duration of antibiotics use within 28 days

The second co-primary outcome (**P2**) is total duration of antibiotics use (in hours) within 28 days from treatment start.

Precisely, total duration is defined as the sum of the durations of each treatment episode, within

28 days. Each duration is computed from the time and day of start and end of treatment[3], registered by the clinicians and nurses. The start time and date are observed and registered accordingly. For end of treatment, what is registered corresponds to the time and day at which the prescribed treatment expires. The corresponding doses are given to the parents before discharge. No data are available on whether the doses were actually received by the newborn, but unlike with some populations of adults patients, we can reasonably expect perfect or nearly perfect adherence/compliance here, with newborns and their caring parents. Note also that the treatment cannot be prolonged by the parents without the consent of the clinicians, since the parents only receive the remainder number of doses, not a bottle.

The second co-primary clinical question of interest is:

"Is the median total duration of antibiotics use within 28 days after the start of treatment lower when using the individualized treatment duration than when using the standard duration of seven days, regardless of initiation of any additional interventions during the follow-up, when needed to ensure good clinical care?"

We consider that both the correpsonding "intention-to-treat" (ITT) and the "Per protocol" (PP) estimands are relevant, as for the first co-primary outcome. Accordingly, the corresponding co-primary estimands are defined as the estimands of Section 4.1, except for the endpoint attributes, which are now defined by **P2** instead of **P1** and the population-level summary attributes, which are now defined by a difference in medians. Although no death is expected, note that here the intercurrent "death" is still addressed by the endpoint: antibiotics use is by definition zero after death ("composite variable strategy", see [15]).

4.2 Secondary endpoint, objectives and estimands

There is only one secondary outcome:

S1: Readmission due to bacterial infection within 100 days from treatment initiation.

It is defined similarly as the first co-primary outcome (**P1**), but within 100 days instead of 28. The corresponding secondary objectives are to estimate the risk of the outcome under each treatment strategy; as well as the risk differences. The corresponding clinical questions of interest are:

"What are the risks of readmission due to bacterial infection within 100 days from treatment initiation when using when using the standard and when using the individualized duration, regardless of initiation of any additional interventions during the follow-up, when needed to ensure good clinical care? What is the corresponding risk difference?"

We consider that both the "intention-to-treat" (ITT) and the "Per protocol" (PP) estimands are relevant, as for the first co-primary outcome **P1** and for similar reasons. Accordingly, the corresponding estimands are defined as for **P1** in Section 4.1, except for the endpoint attributes, which are defined by **S1**.

[3] pricisely, time and day of end of treatment or 28 days after first treatment initiation, whichever comes first (because we compute the sum within 28 days).

4.3 Exploratory outcomes

Exploratory outcomes are listed below, in arbitrary order:

E1: Length of initial hospital stay, from onset of first antibiotic treatment to discharge (within 28 days).

E2: Total length of hospital stay (i.e., cumulative, any stay, within 28 days). For the initial stay, the duration is defined as for **E1**. For subsequent stays, if any, we consider the total duration from admission to discharge.

E3: Culture positive infection within 28 days after initial antibiotic treatment started.[4]

E4: Any serious adverse event (SAE) related to the study intervention within 100 days from treatment start, defined as any of these events:

- Bacteremia
- Meningitis
- Culture positive Sepsis
- Culture positive Meningitis
- Death

E5: Total duration of antibiotics use (in hours) within 100 days from treatment start.

E6: Breastfeeding at 2 days after end of initial antibiotic treatment (accepted range 1-3 days).

E7: same as **E6**, but at 21 days instead of 2 (accepted range 17-25 days).

E8: CRP at 2 days after treatment stopped (accepted range 1-3 days).

E9: same as **E2**, but within 100 days instead of 28.

E10: hospital readmission for any cause within 28 days.[5]

E11: same as **E10**, but within 100 days instead of 28.

Remarks: First, computation of **E5** is similar to that of **P2** detailed in Section 4.1.3, within 100 days for **E5** instead of 28 for **P2**. In Denmark, all drug prescriptions from all doctors are registered in the Shared Medication Record ("Fælles Medicinkort", aka FMK). The necessary data to compute **E5** (and also **P2**) will be extracted from this central registration system. Second, data about **E6** and **E8** are collected during a planned follow-up visit. Third, data about **E7** are collected by phone interviews.

[4] **Note:** about **E3**, we mean a new culture started after inclusion. If the index culture that was negative at inclusion turns positive later, this does not count.

[5] Unlike for the primary outcome, there is no restriction about the admission to define **E10**, except for the admission to be within 28 days.

5 General Considerations

5.1 Statistical Hypotheses

The following (confirmatory) 1-sided hypothesis is planned to be tested, in relation to the first co-primary objective (detailed in Sec. 4.1). The aim is to show that "individualized" duration is not unacceptably worse than "standard" duration with respect to readmission due to bacterial infection within 28 days, using a non-inferiority margin of 4%. The aim is to show this using the two estimands: the "intention-to-treat" estimand and "Per Protocol" estimand. Importantly, non-inferiority will only be concluded if both are significant.

- Null hypothesis: the risk of readmission due to infection within 28 days from start of treatment is **at least** 4% higher when using the "individualized" treatment duration than when using the "standard" treatment duration. Formally, $\mathcal{H}_0 : \pi_1 - \pi_0 \geq 4\%$, where π_1 and π_0 are the 28-day risks using "individualized" and "standard" duration, respectively.

versus

- Alternative hypothesis: the risk of readmission due to infection within 28 days is **at most** 4% higher when using the "individualized" treatment duration than when using the "standard" duration. Formally, $\mathcal{H}_1 : \pi_1 - \pi_0 < 5\%$.

The following (confirmatory) 1-sided hypothesis is planned to be tested, in relation to the second co-primary objective (detailed in Sec. 4.1). The aim is to show that "individualized" is better than "standard" with respect to the median of the total duration of antibiotics use within 28 days from treatment start. The aim is to show this using the two estimands: the "intention-to-treat" estimand and "Per Protocol" estimand. Importantly, superiority will only be concluded if both are significant.

- Null hypothesis: the median of the total duration of antibiotics use within 28 days from start of treatment is **larger** when using the "individualized" treatment duration than when using the "standard" treatment duration. Formally, $\mathcal{H}_0 : m_1 > m_0$, where m_1 and m_0 are the median total duration using "individualized" and "standard" initial treatment duration, respectively.

versus

- Alternative hypothesis: the median of the total duration of antibiotics use within 28 days is **smaller** when using the "individualized" treatment duration than when using the "standard" duration. Formally, $\mathcal{H}_1 : m_1 < m_0$.

Operationally, the hypotheses will be evaluated by one-sided tests at 2.5% and matching 95% two-sided confidence intervals.

5.2 Multiplicity Adjustment

No multiple testing correction will be used, as formal hypothesis testing will be performed only for the primary estimands described in Section 4.1. Although there are two co-primary outcomes with two estimands each and hence four hypothesis tests, no multiple correction is needed. This is

because we require that the two co-primary objectives of the trial have been met to conclude that the trial is positive. This corresponds to an "All-or-none" win criterion [6] and "intersection union test" [1, Sec. 2.2.2]. In addition, we require that both tests, for the ITT and PP estimands, are significant to conclude statistical significance for the corresponding co-primary objective.

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 [20].

5.3 Missing Data

No missing data are expected for co-primary outcome **P1**, as the investigators have access to all relevant data from all hospital admissions (and deaths) occurring in Denmark. In Denmark, these data are registered centrally. The same naturally holds for **S1**, **E10** and **E11**.

No missing data is anticipated either for co-primary outcome **P2**. As already mentioned in Section 4.3, in Denmark, all drug prescriptions from all doctors are registered in the Shared Medication Record ("Fælles Medicinkort", aka FMK). The investigators will extract the necessary data to compute **P2** (total duration of treatment within 28 days) and **E5** (total duration of treatment within 100 days) from this central registration system. Hence, there will be no missing data.

No missing data is expected for length of hospital stay (neither total length, **E2**, **E9**, nor initial length, **E1**), culture positive infections (**E3**) or SAE (**E4**). However, some missing data are expected for data about breastfeeding (**E6** & **E7**). We might also observe a few missing data for the exploratory outcome CRP at 2 days after treatment stopped (**E8**). The most probable cause of missing data is omission of registration by the treating physician as all infants should be seen and have CRP level checked at that time. More missing data are expected for breasfeeding (**E6**) than for CRP at 2 days (**E8**). This is because data collection about breasfeeding needs the treating clinicains to write down the information, whereas data collection about CRP at 2 days is automated. The assumption of the data being missing completely at random is considered realistic for exploratory outcome **E6** and **E8**. Data collection for **E7** was planned via phone interviews. Missing data completely at random is also considered realistic for **E7**.

Note that data about breasfeeding and CRP outcomes collected at follow-up visits (**E6**,**E7**,**E8**) will be considered as missing if the visit occurs outside of the accepted range (1-3 days for the "visit at 2 days"; 17-25 days for the "visit at 21 days"). However, no visit or very few are expected to occur outside the accepted ranges.

5.4 Covariate adjustment

No covariate adjustment will be used in the statistical analysis. Non-parametric unadjusted analyses will be performed instead.

Because randomization was stratified by maximum CRP, covariate adjustment would usually be recommended, especially for the analysis of the co-primary outcomes **P1** and **P2**, as this usually leads to power gains [10, 16]. However, we expect very few hospital readmissions due to bacterial infection (**P1**) in each group, whatever the maximum CRP. This specific context has two important consequences for the analysis of outcome **P1**. First, potential loss of power is expected to be negligible. Second, the typical large sample approximations required to ensure correct coverage probability and type-I error control using a "covariate-adjusted" approach would be very questionable.

Covariate adjustment could also be recommended to make the "no unmeasured confounding" assumption more plausible, for the analysis of the PP estimand using the "Per protocol analysis set" (defined in Section 6). Here again, the rationale for no covariate adjustment is that we expect very few hospital readmissions due to bacterial infection (**P1**), making "covariate-adjusted" inference likely unreliable. Instead, a sensitivity analysis will be provided (see Section 7.1.1). Additionally, descriptive statistics will be provided of newborns included in, and excluded from, the "Per protocol analysis set". This will facilitate the assessment of the risk of selection bias, which could potentially lead to confounding (see Section 10.2.5).

A very large power is expected for the analysis of the co-primary outcome **P2**, hence there is no real power gain incentive to use covariate adjustment in the analysis of **P2**. Additionally, we are not aware of well-known, model-robust and recommended methods using covariate adjustment to compare medians. Finally, the analysis of **P2** aims to establish superiority and unadjusted ITT analyses are generally conservative when the aim is to show superiority.

6 Analysis Sets

- The **"All participants analysis set" (All)** consists of all randomized participants.
- The **"Per protocol analysis set" (PP)** consists of all randomized participants that have actually received the intervention ("individualized" or "standard") as they should according to the protocol. Deviations form the protocol that are used to exclude patients from this analysis sets are listed in Section 10.2.5.
- The **"Breastfeeding-2d analysis set" (B2)** consists of all randomized participants that do not have missing data about breastfeeding at 2 days after end of initial antibiotic treatment (i.e., no missing data for **E6**).[6]
- The **"Breastfeeding-21d analysis set" (B21)** consists of all randomized participants that do not have missing data about breastfeeding at 21 days after end of initial antibiotic treatment (i.e., no missing data for **E7**).[6]
- The **"CRP-2-ITT analysis set" (CRP2)** consists of all randomized participants that do not have missing data about CRP at 2 days after treatment stopped (i.e., no missing data for **E8**).[6]
- The **"Breastfeeding-2d-PP analysis set" (B2-PP)** consists of all randomized participants that are in both the "Per protocol analysis set" and the "Breastfeeding-2d-ITT analysis set".
- The **"Breastfeeding-21d-PP analysis set" (B21-PP)** consists of all randomized participants that are in both the "Per protocol analysis set" and the "Breastfeeding-21d-ITT analysis set".
- The **"CRP-2-PP analysis set" (CRP2-PP)** consists of all randomized participants that are in both the "Per protocol analysis set" and the "CRP-2-ITT analysis set".

Note: abbreviations, in parenthesis are used in Table 2 below.

[6] Note that data collected during a visit or phone interview occuring outside of the accepted range will be conisdered as missing, as already stated in Section 5.3. Accepted ranges for visit times are 1-3 days for visits related to outcomes **E6** and **E8** and 17-25 days for phone interviews related to outcome **E7**.

7 Analyses supporting the primary objectives

The main analyses for the ITT estimands will use the "All participants analysis set" and corresponds to "intention-to-treat" analyses. The main analyses for the PP estimands will use the "Per protocol analysis set" and corresponds to "per protocol analysis" analyses. Apart from the difference in analysis set being used, the PP and ITT analyses will be performed identically, as described below.

7.1 Analyses of co-primary outcome P1

We will not adjust for baseline covariates in the computation of the CIs and p-values. Specifically, we will compute the two-sided 95%-CI for the risk difference using the Miettinen-Nurminen asymptotic score interval method [19], as it has been shown to perform well and to be *"safe to use"* by Fagerland et al. [9], even when few events are observed. A matching p-value will be computed too. To compute the CI, we will use the `diffscoreci()` function from the `PropCIs` package of `R`, as suggested by Fagerland et al. [9]. To compute the p-value, we will use the `z2stat()` function of the same package; see details in Appendix A.1. This method is more appropriate than simpler (common) large sample approximations when a small to moderate number of events is expected, but it is not overly conservative [9]. The risk and risk difference estimates will be computed as the empirical (i.e., observed) proportions and their difference. Two-sided 95%-CI for the risks in each arm will be computed as exact binomial two-sided 95% CIs, using the `binom.test()` function of `R`.

For completeness and especially to report the timing of the readmissions, cumulative incidence plots will also be reported, as described in Section 10.1.4. Further details about each admission will also be reported, as described in Section 10.1.1.

7.1.1 Sensitivity analysis

The analysis for the PP estimand might be biased if the newborns excluded from the "Per protocol analysis set" differ from those included in it. Hence, for completeness and to assess the robustness of the results to a possible selection bias, we will perform the following sensitivity analysis.

We will report the results (p-value and upper limit of the 95% CI) assuming that zero, one, two, three, ... of the newborns excluded from the "Per protocol analysis set" in each arm have experienced the co-primary outcome **P1**. A graphical representation similar to the "Enhanced tipping-point display" presented in Liublinska et al. [18, Fig. 2] will be reported, to provide an overview of the results in these cases. It will show a heat map of the upper limit of the 95% CI (or p-value) for the different cases.

7.2 Analyses of co-primary outcome P2

The analysis of **P2** will consist of a simple unadjusted analysis to compare medians. We will estimate the median for each arm and their difference, together with two-sided 95%-CIs. Point estimates will be computed as sample medians (via the `median()` function of `R`) and standard errors by standard non-parametric bootstrap (using 10,000 bootstrap samples). The 95% CIs will be computed as "Est. $\pm$ 1.96·SE", where "Est." denotes the point estimate and SE the corresponding standard error. Similarly, the p-value for the one-sided test will be computed using the test statistic Z=Est./SE, assuming that Z follows a standard normal distribution.

To complement the above analysis and for the sake of completeness, a dot plot or histogram will be reported to summarize the distribution in each arm. Quartiles of the distributions in each

arm will be reported too (e.g., boxplots will be overlaid on top of the plot). We will also consider showing on the plot the observations corresponding to the newborns who experienced the primary outcome.

8 Analyses supporting secondary objectives

The secondary outcome **S1** and corresponding estimands are similar to those of the co-primary outcome **P1**. Therefore, it will be analyzed similarly, as described in Section 7, except that no p-value will be reported (as already mentioned in Section 5.2).

9 Analyses of exploratory outcomes

The table 2 lists the analysis sets and methodology that will be used to analyze each exploratory outcome. The method will be either that used for analyzing **P2** described in Section 7.2 (i.e., non-parametric bootstrap comparison of medians, for quantitative outcomes) or that used for **S1** described in Section 8 (Miettinen-Nurminen, for binary outcomes). Note that two analysis sets are listed for each analysis, indicating that the analysis will be performed twice, using each set, to perform both PP and ITT analyses.

Outcome	Analysis set(s)	Method as for analyzing
E1	All, PP	P2
E2	All, PP	P2
E3	All, PP	S1
E4	All, PP	S1
E5	All, PP	P2
E6	B2, B2-PP	S1
E7	B21, B21-PP	S1
E8	CRP2, CRP2-PP	P2
E9	All, PP	P2
E10	All, PP	S1
E11	All, PP	S1

Table 2: Analysis of exploratory outcomes

Remarks:

(i) Exploratory outcomes **E6** and **E7** are not binary outcomes. They are categorical outcomes, with possible values "no breastfeeding", "partial breastfeeding" or "exclusive breastfeeding". Hence, we will report the proportions and the between arm differences in proportions for each of these modalities. Additionally, we will also perform the analysis after dichotomizing into "no breastfeeding" vs "partial or exclusive breastfeeding".

(ii) The pre-specified analysis of **E6**, **E7** and **E8** are complete case analysis. For completeness and transparency, descriptive statistics will be provided, per arm, about baseline characteristics of subjects with a missing value. Any difference to subjects without missing data thought as clinically relevant will be reported.

10 Other analyses

10.1 Additional outcome analyses

10.1.1 Details about readmissions

Few readmissions are expected. Hence, it will be possible to provide details about each readmission (including admissions which do not meet the definition of the primary outcome). This will be done with the main aim to detail the cause of each readmission and what happened during admission and after discharge, as relevant.

10.1.2 Culture positive infections

For completeness, the types of infection will be provided, for each positive culture, to complement the main results about exploratory outcome **E3**, per arm. The data will be summarizes descriptively (via counts and proportions).

10.1.3 SAE

To complement the main analysis of **E4** (SAE), we will also compute the risk and risk difference, with 95%-CI, for each possible SAE listed in the definition of SAE. The same statistical methods as those used for the analysis of **E4** (SAE) will be used.

10.1.4 Cumulative incidence plots

Cumulative incidence curves will be reported, per arm, to describe when each of the following events occurred during the follow-up.

(I) readmission due to bacterial infection, within 100 days (relevant for **P1** and **S1**)

(II) culture positive infection within 28 days (**E3**)

(III) SAE within 100 days (**E4**)

The curves will be estimated by simple empirical (i.e., observed) proportions and pointwise two-sided 95%-CI computed as exact binomial two-sided 95% CIs (computed using the `binom.test()` function of `R`).

We will perform the analysis using the "All participants analysis set". We might reproduce the analysis using the "Per protocol analysis set", if relevant, although we do not expect it will be the case, as we expect very few non adherers.

10.1.5 Subgroup analyses

We plan to perform subgroup analyses for these subgroups:

- newborns with maximum CRP level within the last 48h before randomization $>$ 90 mg/l
- newborns with maximum CRP level within the last 48h before randomization $\leq$ 90 mg/l

(note that randomization was stratified using the two subgroups above)

- "High risk" newborns, defined as newborns with onset of treatment $<$ 18 hours after birth, and duration of symptoms $\geq$ 24 hours.
- "Low risk" newborns, defined as onset of treatment $\geq$ 18 hours after birth and duration of symptoms $<$ 24 hours.
- "Medium risk" newborns, defined as newborns neither at "High risk" nor at "Low risk".

For each subgroup, we will replicate the analyses for the two co-primary outcomes and the secondary outcome, described in Sections 7 and 8.

10.2 Descriptive analyses

10.2.1 Recruitment

Recruitment of the infants 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 [22].

10.2.2 Screening data

Screening data, about assessment for eligibility, were collected. They will be presented in a in a flow diagram, inspired by the CONSORT guidelines [22].

10.2.3 Baseline characteristics

Baseline characteristics will be descriptively summarized per randomized arm, using the "All participants analysis set". The list of baseline variables to be summarized includes:

Maternal characteristics

- Maternal age (years)
- Parity (0, 1, 2 or $\geq$3)
- Maternal risk factors of infection (one or more, yes/no)
  - Prolonged rupture of membranes ($>$ 18 hours) (yes/no)
  - Meconium stained amniotic fluid (yes/no)
  - Maternal fever ($> 38.0^{\circ}$C) (yes/no)
  - Intrapartum antibiotics (yes/no)
  - Suspicion of chorioamnionitis (yes/no)
  - Group B Streptococcus status (yes/no)
    - Previous pregnancy: Colonized or child born with GBS disease (yes/no)
    - This pregnancy: Colonized (yes/no)
  - Chorioamnionitis (yes/no)

Newborn characteristics

- Sex (male/female)
- Gestational age (weeks and days)
- Birth weight (gram)
- Apgar Score at 5 minutes (range 0-10)
- Mode of delivery (Vaginal/Instrumental/Elective C-section/Non-elective C-section)
- Symptoms at onset of antibiotics (one or more, yes/no)
  - Behavioral changes (yes/no)
  - Tachypnea ($>$ 60 breaths/minute) (yes/no)
  - Heart rate at onset $>$ 180 beats/minute (yes/no)
  - Capillary refill time $>$ 3 seconds (yes/no)
  - Temperature $< 36.5°$C or $> 38.0°$C (yes/no)
- Supportive therapy (one or more, yes/no)
  - CPAP or nasal High Flow (yes/no)
  - Mechanical ventilation (yes/no)
- Age at onset of antibiotic treatment (hours)
- Initial treatment with benzyl-penicillin and gentamicin (yes/no)
- maximum CRP within 48h before randomization (mg/L)
  - Above 90 mg/L (yes/no)
  - Elevated ($>$ 10 mg/L) (yes/no)
- Blood culture taken (yes/no)
- Lumbar puncture taken (yes/no)

Other characteristics

- Study site

For quantitative variables, we will present median, first and third quartiles and also minimum and maximum. Categorical variables will be summarized by counts and percentages. The number and proportions of missing values (if any) will be reported for each variable, per arm. 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 [21].

10.2.4 Maximum CRP

We will report descriptive statistics, specifically median, first and third quartiles, minimum and maximum, per arm, about:

- maximum CRP observed during index admission.
- age (in hours) at this maximum

For compleness, we will also report the proportions of newborns for whom maximum CRP during index admission was not observed within 48h before randomization (none or very few expected).

10.2.5 Adherence

We will document the number of newborns who did not receive the treatment strategy to which they were randomized, per arm. The reasons will be provided too, whenever available. These newborns are those excluded from the "All participants analysis set" to define the "Per protocol analysis set" (see Section 6).

Daily data about both antibiotic treatment and symptoms were collected, thus it is possible to identify these newborns. Specifically, we will report the numbers and proportions of:

- newborns randomized to "individualized" duration who continued antibiotics longer than they should according to the "individualized" duration treatment strategy.
- newborns randomized to "standard" who stopped antibiotic treatment before they should according to the "standard" duration treatment strategy.

Note that an inconsistency of one day (or less) between the treatment duration observed and the duration that should be observed according to the randomized treatment strategy will not be considered as an inconsistency. That is, these newborns will not be counted in the above proportions. This is because, among other things, decisions to stop or continue the treatment can happen at any time of the day.

Descriptive statistics of newborns included and excluded from the "Per protocol analysis set" will be provided, per arm, to assess the risk of selection bias possibly leading to unmeasured confounding. Specifically, similar descriptive statistics of baseline characteristics as those described in Section 10.2.3 will be provided. Any noticeable difference between newborns included in the PP analysis set and those excluded will be noted, whenever though as possibly clinically relevant.

10.2.6 Oral and intravenous antibiotic treatment

For each arm, we will report how many newborns switch from intravenous to oral antibiotic treatment and when. For completeness and to document the timing of the switches, cumulative incidence plots, per arm, will be provided. We will also report the median (if more than half indeed switch) or another quantile of the time to switch (if less than half switch), per arm, for completeness. It is expected that most switches happen close to 48 hours after treatment started.

10.2.7 Symptoms disappearance

Daily data about symptoms were registered. Cumulative incidence plots of symptoms disappearance, per arm, will be provided. Median times will also be reported and compared. A 95%-CI for the difference in median will also be provided (computed as detailed in Section 9)

10.2.8 Visits at 2 days after end of treatment and phone interviews at 21 days

We will descriptively summarize the timing of:

- the visit planned 2 days after end of treatment.
- the phone interview planned 21 days after end of treatment.

This is during this visit that data about exploratory outcome **E6** and **E8** are collected; and during the phone interview that data about exploratory outcome **E7** are collected. Some visits might happen earlier, later or not at all. Median, first and third quartiles and also minimum and maximum will be reported, per arm.

11 Sample size determination and power calculations

11.1 Sample size determination

The sample size calculation was performed to properly power the analysis of the first co-primary endpoint (**P1**, readmission). The power of the analysis for the second co-primary outcome (**P2**, total duration) was expected to be considerably larger. Hence, the impact of the analysis of **P2** was considered negligible in the power and sample size calculations for this study.

A sample size of $n = 488$ patients (244 in each group) was calculated as follows. The desired power was 90% and the type one error was set to 2.5% (one sided test). The expected proportion of the first co-primary endpoint (readmission) was 1% in the control group and 1.5% in the experimental group. These proportions were expected based on Danish data about births in 2018-2020 (unpublished at the time of planning the trial, but now in [3]). They aligned with expected treatment success in other studies examining other aspects of early onset infection trial [17].

Using a non inferiority margin of 4% for the difference in proportions and the usual asymptotic normal approximation gives

$$\text{Power} \approx \Phi\left(\frac{0.04 - (0.015 - 0.01)}{\sqrt{0.015(1-0.015)/207 + 0.01(1-0.01)/207}} - 1.96\right) \approx 90\% \ ,$$

where Φ denotes the cumulative standard normal distribution function; see e.g. Chow et al. [4, Sec. 4.2.2]. This suggests a sample size of size of $n/2$ =207 per group, but the sample size was further rounded up to 244=207/(1-0.15) per group, so $n = 488$ in total, to account for the possibility that up to 15% of the newborns might not be included in the Per Protocol analysis set.

11.2 Additional power calculations

Because the expected number of events in each arm are very low ($\approx 1.5\% \times 244 = 3.6$ and $\approx 1\% \times 244 = 2.4$), it has been realized (at time of SAP writing) that the initial asymptotic power calculations (presented above) were probably not sufficiently precise. Hence, additional (nearly exact) power computations were performed. The main results are provided in Figures 1 and 2. They suggest that the study is substantially underpowered if the risks in each arm are indeed 1% and 1.5%, but decently powered if the risks are lower, which is not unrealistic. Specifically, the power is computed as 67.3% instead of 90% under the assumptions used in the initial power calculation, as shown in 1.

As a side remark, similar computation suggested that the risk of type-I error is controlled exactly for any risk below 6%, even though an asymptotic method is used (Miettinen-Nurminen asymptotic score). This is in line with the conclusions of Fagerland et al. [9] already mentioned in Section 7.

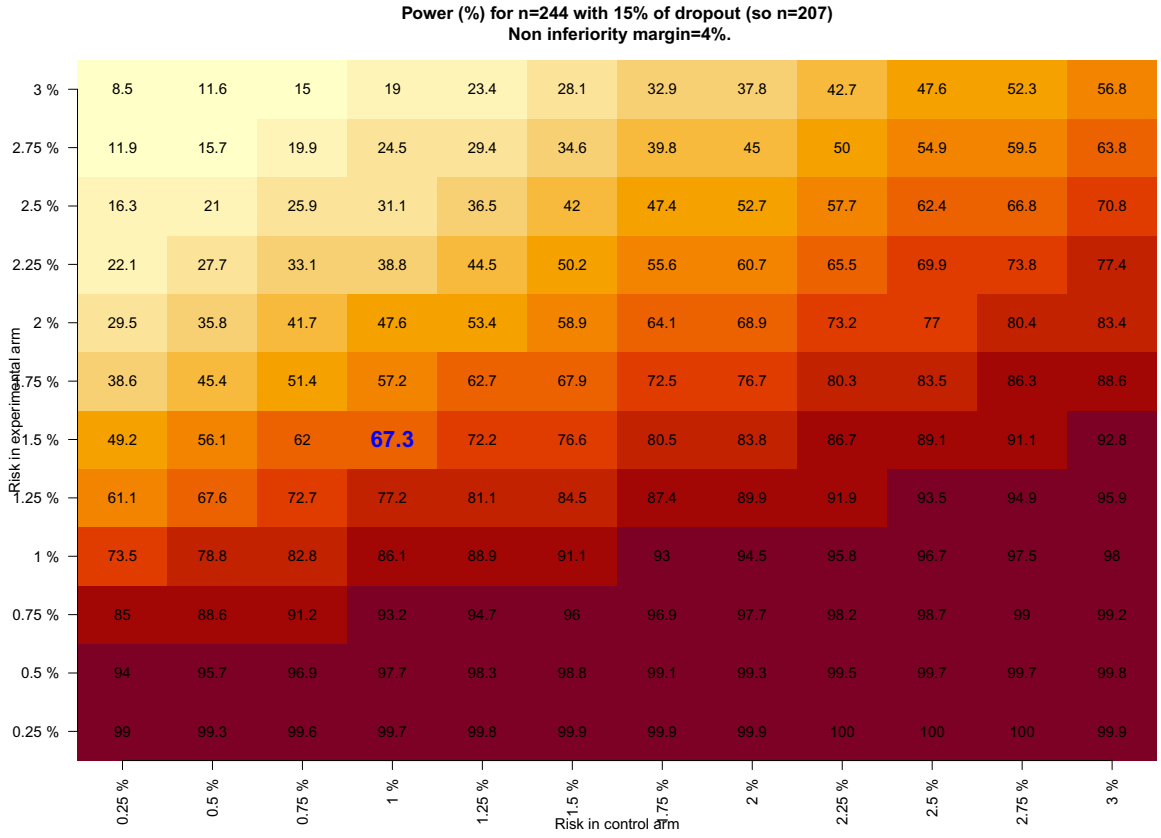

Figure 1: Additional power computation (PP analysis, n=207)

12 Changes to protocol-planned Analyses

1. The definition of the first co-primary outcome (**P1**) was slightly changed. It is now defined as readmission within 28 days from treatment start instead of 21 after treatment stopped. This is a very minor change as all or nearly all readmissions are expected to happen soon after after end of treatment. This is a **conservative change**, because the "individualized" treatment strategy lasts between 2 and 7 days, which is less long than the "standard" treatment strategy (7 days). Note that 7 (standard duration)+21=28 days. Hence this change can only lead to more events observed in the experimental arm and thus a larger risk difference making it more difficult to reach non-inferiority. The decrease in power due to this change is expected to be negligible, as again all or nearly all readmissions are expected to occur soon after after end of treatment. The rationale for this change is that it makes the definition of the primary outcome independent of the treatment actually received, which complies with the usual aim to make the two arms as similar as possible with respect to everthing except the treatment startegies being investigated and compared. This usal aim, among other things,

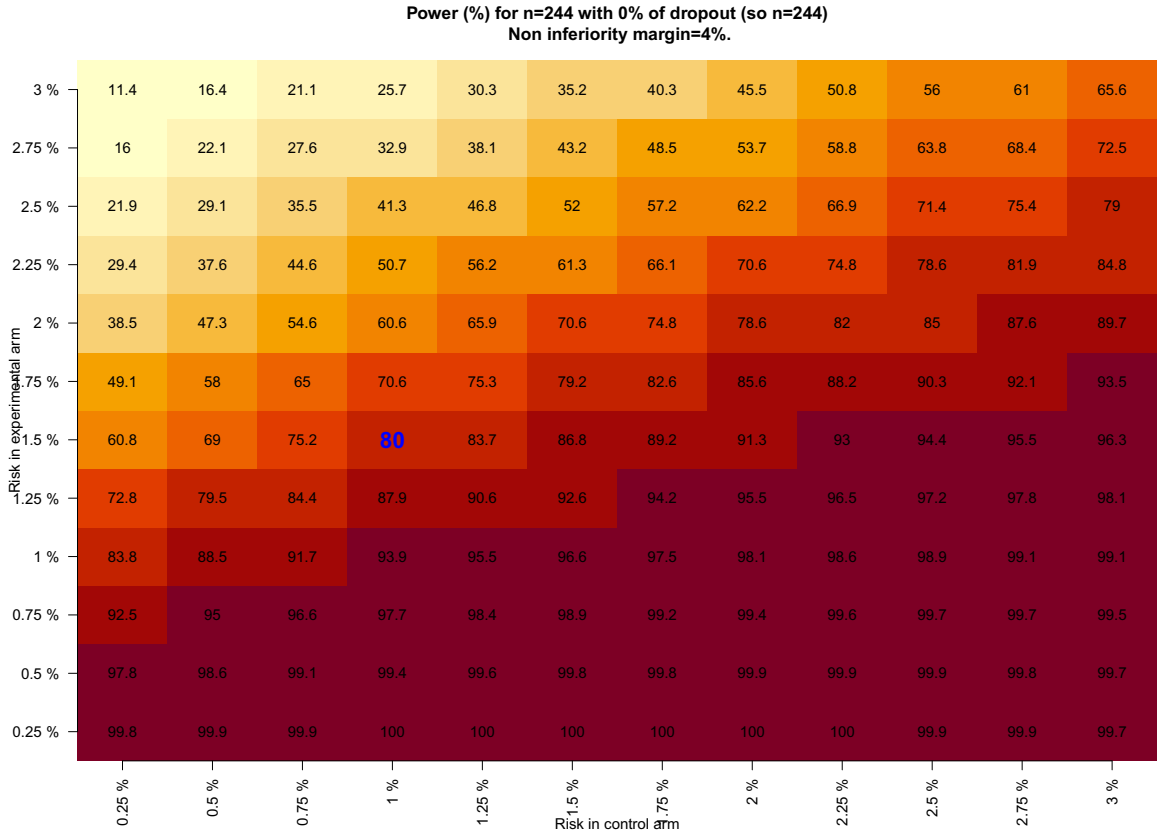

Figure 2: Additional power computation (ITT analysis, n=244)

is the main motivation for randomization. With the original definition of the outcome, even in the absence of treatment effect, a difference (anti-conservative bias) could occur if e.g., children were more exposed to pathogens and hence more likely to be infected during their fourth week than during the three first weeks. We cannot rule out for sure that this is not the case due to e.g., the newborn family having more social activities with friends or extended family as the newborn gets older. Note that the timing of the readmissions will be provided (see Section 10.1.4), such that the potential impact of this change on the number of observed primary events will be transparently reported.
Also, in this SAP, more details are provided to define the first co-primary outcome of readmission due to bacterial infection (**P1**). It was defined was originally described as "Readmission due to infection, defined as symptoms, affected biomarkers and antibiotic treatment $>$ 72 hours". In Section 4.1 of this SAP, symptoms are listed and "affected biomarkers" is more specifically described as maximum CRP during readmission $>$ 10 mg/l.

2. Similarly and for the same reason, the exploratory outcome **E3** is now defined within 28 days from treatment start instead of within 21 after treatment stopped.

3. CRP at 2 days after treatment stopped is now considered as an exploratory outcome (**E8**). The reason for this change is that it has never been the aim to present "strong" evidence of anything specific for that outcome, but just to document what can be expected following

each treatment strategy.

4. Exploratory outcomes **E10** and **E11** have been added, for completeness.

5. Note that it has been unintentionally registered at ClinicalTrials.gov that death was a co-primary outcome. What was meant was that the primary outcome is a composite outcome of readmission and death, as it is writen in this SAP (but we do not expect any death).

References

[1] Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple comparisons using R*. Chapman & Hall/CRC.

[2] Brittain, E. and Lin, D. (2005). A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Statistics in Medicine*, 24(1):1–10.

[3] Carlsen, E. L. M., Dungu, K. H. S., Lewis, A., Vissing, N. H., Aunsholt, L., Trautner, S., Stanchev, H., Dayani, G. K., Pedersen, A.-J. L., Bjerager, M., et al. (2024). Switch from intravenous-to-oral antibiotics in neonatal probable and proven early-onset infection: a prospective population-based real-life multicentre cohort study. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 109(1):34–40.

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

[5] CMP (2000). Points to consider on switching between superiority and non-inferiority. Technical report, EMA, `https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-and-non-inferiority_en.pdf`.

[6] Dmitrienko, A., D'Agostino Sr, R. B., and Huque, M. F. (2013). Key multiplicity issues in clinical drug development. *Statistics in medicine*, 32(7):1079–1111.

[7] Ekelund, K. and Konradsen, H. (2004). Invasive group b streptococcal disease in infants: a 19-year nationwide study. serotype distribution, incidence and recurrent infection. *Epidemiology & Infection*, 132(6):1083–1090.

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

[9] Fagerland, M. W., Lydersen, S., and Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, 24(2):224–254.

[10] 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.

[11] Fleming, T. R. (2008). Current issues in non-inferiority trials. *Statistics in Medicine*, 27(3):317–332.

[12] 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.

[13] Hernán, M. A. and Hernández-Díaz, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clinical trials*, 9(1):48–55.

[14] ICH E9 (1998). Statistical principles for clinical trials. Technical report, EMA/CHMP/ICH, www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

[15] 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.

[16] 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.

[17] Keij, F. M., Kornelisse, R. F., Hartwig, N. G., Mauff, K., Poley, M. J., Allegaert, K., Reiss, I. K., and Tramper-Stranders, G. A. (2019). Rain study: a protocol for a randomised controlled trial evaluating efficacy, safety and cost-effectiveness of intravenous-to-oral antibiotic switch therapy in neonates with a probable bacterial infection. *BMJ open*, 9(7):e026688.

[18] 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.

[19] Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4(2):213–226.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Stocker, M., Van Herk, W., El Helou, S., Dutta, S., Fontana, M. S., Schuerman, F. A., van den Tooren-de, R. K., Wieringa, J. W., Janota, J., van der Meer-Kappelle, L. H., et al. (2017). Procalcitonin-guided decision making for duration of antibiotic therapy in neonates with suspected early-onset sepsis: a multicentre, randomised controlled trial (neopins). *The Lancet*, 390(10097):871–881.

[25] TransCelerate (2024). Common Statistical Analysis Plan Template v5. https://www.transceleratebiopharmainc.com/assets/clinical-content-reuse-solutions/. Accessed: 2024-05-01.

[26] Zangwill, K. M., Schuchat, A., and Wenger, J. D. (1992). Group b streptococcal disease in the united states, 1990: report from a multistate active surveillance system. *MORBIDITY AND MORTALITY WEEKLY REPORT: CDC Surveillance Summaries*, pages 25–32.

A Appendix

A.1 R code for the main analysis of P1

```
# x1: number of events (readmission) observed in the experimental group
# n1: sample size of the experimental group
# x2: number of events (readmission) observed in the control group
# n2: sample size of the control group
library(PropCIs)
resCI <- diffscoreci(x1=x1,n1=n1,x2=x2,
                     n2=n2,conf.level=0.95) # two-sided 95%-CI of risk difference
StatTest <- z2stat(x1/n1, n1, x2/n2, n2, 0.04)  # matching Chi-2 test-statistic (NI margin=4%)
results <- c(risk1=x1/n1,                 # risk in experimental group
             risk2=x2/n2,                 # risk in the control group
             difference=x1/n1-x2/n2,      # difference
             lower=resCI$conf.int[1], # lower bound of two-sided 95%-CI of difference
             upper=resCI$conf.int[2], # upper bound of two-sided 95%-CI of difference
             pvalue=pnorm(sign(x1/n1-x2/n2-0.04)*sqrt(StatTest),
                          lower.tail=TRUE)) # one-sided p-value (NI margin=4%)
results
```