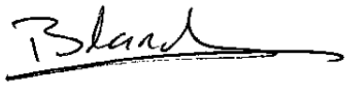
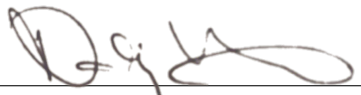
Statistical Analysis Plan (SAP)

SAP version number:	2025–01–13
Person writing the SAP:	Paul Blanche
Principal investigators:	Nadja Hawwa Vissing & Ulrikka Nygaard
Protocol title:	Early termination of empirical antibiotics in febrile neutropenia in children with cancer
ClinicalTrials.gov ID:	NCT04637464

Note: inspiration to write this SAP came from the SAP template provided by TransCelerate [25] as well as recommendations from Gamble et al. [11], Stevens et al. [23] and Evans and Ting [6].

Contents

1 Statistical analysis plan approval signature page

Person writing the SAP:	Name: Paul Blanche
	Position/Title: Associate Professor in Biostatistics
	 January 13, 2025
	Signature & Date
Approved by:	Name: Nadja Hawwa Vissing
	Position/Title: MD PhD, Senior Consultant
	 January 15, 2025
	Signature & Date
	Name: Ulrikka Nygaard
	Position/Title: MD, PhD, Associate Professor, Senior Consultant
	Ulrikka Nygaard 15-01-2025
	Signature & Date

2 List of Abbreviations

- ALL: Acute Lymphoblastic Leukemia
- AML: Acute Myeloid Leukemia
- ANC: Absolute Neutrophil Count
- ANCOVA: Analysis of covariance
- CI: Confidence Interval
- GEE: Generalized Estimating Equations
- ITT: Intention-To-Treat
- SAE: Severe 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, clinical trial of "short" versus "standard" (i.e., "long") empirical antibiotics treatment duration of febrile neutropenia in high-risk children with cancer. Included children have an absolute neutrophil count (ANC) $< 0.5 \times 10^9$cells/L and an expected duration of neutropenia of more than seven days. The primary endpoint is the number of days without antibiotic treatment within the first 28 days after treatment initiation. It is hypothesized that a "short" empirical antibiotics treatment duration is sufficient to prevent reinfection that would require additional antibiotic treatment days, hence superior to "standard" treatment duration. Randomization was stratified by site, via block randomization with random block size. Patient accrual has stopped early because of lower accrual rate than anticipated, as detailed in Sec. 3.1.

This SAP has been written after this decision to stop the trial early was taken. 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.

3.1 Early termination of the trial

The study had a lower accrual rate than anticipated, presumably for two reasons. First, the COVID-19 lockdown and restrictions in Denmark led to fewer leukemia diagnoses and fewer febrile episodes among the patients. Second, changes in standard cancer treatment protocols resulted in shorter neutropenia periods. This resulted in fewer patients fulfilling the inclusion criteria at the study sites.

Initially, the study period was extended by one year within the available funds (2020–2022 $\rightarrow$ 2020–2023). On November 8th, 2023, a meeting with the study board (BKA, HH, NV, UN, TN) was held. By that time, around 80 episodes had been included. It was decided to extend the study by another year, continuing through 2024, with recruitment ending on December 31st, 2024. Although the target of 220 episodes would not be reached, the sample size was estimated sufficiently large to provide valuable findings. Additionally, an opportunity to pool the data of this trial with

similar data from an Australian study group came up. It is anticipated that a meta-analysis of the two trials will be performed and that it will mitigate the limitations of the reduced sample size. Extending the study to reach the initially planned 220 episodes was estimated to require another five years, which would have led to funding difficulties. Additionally, the decline in the inclusion rate reflects a change in the patient population over time and this would have complicated the interpretation of the results, had the study been further extended.

4 Objectives, Endpoints, and Estimands

4.1 Primary objective, endpoint and estimand

The two treatment strategies (interventions) being investigated are the following.

- **Experimental ("short")**: discontinuation of antibiotics, despite neutrophil count $< 0.5 \times 10^9$ cells/L, after 48 hours of apyrexia and clinical stability. A child is considered clinically stable by the treating physician when there is resolution of all symptoms and signs of infection, and normalization of vital signs including heart rate, respiratory rate, oxygen saturation, blood pressure, and daily diuresis.
- **Control ("standard" or "long")**: discontinuation of antibiotics when either i) neutrophil count is $\geq 0.5 \times 10^9$ cells/L, and the child is afebrile and clinical stable or ii) the child has received 10 days of antibiotics and has been afebrile and clinically stable for 7 days.

The two treatment strategies are followed from randomization until neutrophil recovery (i.e. neutrophil count $\geq 0.5 \times 10^9$) or 28 days, whichever comes last. In case of relapsing fever, the patients continue to be treated according to the same treatment strategy ("short" or "long"). Randomization could occur from 36 hours after initiation of antibiotics until 48 hours of apyrexia and clinical stability (when the two treatment strategies can start to differ).

The primary outcome is the number of days without antibiotic treatment within the first 28 days after treatment initiation. For each of the 28 days, the information of whether the patient received antibiotics has been collected. The typical reasons for restarting antibiotics are new onset of fever, typically:

- febrile neutropenia (without documented infection)
- fever with or without neutropenia due to a possible or proven bacterial infection, including bacteremia (bacterial infection in the blood stream), pneumonia, skin infection, gastroenteritis or typhilitis (infection in the gut), urinary infection or other types of focal infections; side effects to chemotherapy that is accompanied by fever, such as pancreatitis

Restarting antibiotics might also happen without new onset of fever, for instance in case of suspicion of infection (e.g. focal symptoms from skin, lungs, gastrointestinal, etc).

The primary clinical question of interest is: *"Is the mean number of days alive without antibiotic treatment, within 28 days after treatment started, larger when children receive the "short" treatment duration strategy than when they receive the "long", among children with cancer who started antibiotic treatment for high risk neutropenic fever of unknown origin, regardless of any intervention*

during the follow-up needed to ensure good clinical care?

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

- Population: children (aged 0-17) with cancer receiving antibiotic treatment for high risk neutropenic fever of unknown origin 48 hours after presentation. In particular, children with
  - a single temperature of at least 38.5° C , or a temperature above 38.0° C sustained over a 1-hour period (auricular, oral or rectal), at presentation.
  - an absolute neutrophil count $< 0.5 \times 10^9$cells/L and an expected duration of neutropenia of more than 7 days at presentation.
  - a negative blood culture (or other clinically relevant culture) 48h after the start of the culture.
- Endpoint: number of days alive without antibiotic treatment within the first 28 days after treatment initiation.
- Treatment: The investigational interventions ("short" vs "long", as defined above) regardless of adherence and of any subsequent treatment decision or intervention during the follow-up needed to ensure good clinical care ("treatment policy strategy", see [14]).
- The intercurrent event "any subsequent treatment decision or intervention needed to ensure good clinical care" is addressed by the treatment condition of interest attribute ("treatment policy strategy", see [14]). The intercurrent event adherence to the investigational intervention is also addressed by the treatment condition of interest attribute ("treatment policy strategy", see [14]). There are no other relevant intercurrent events anticipated.
- Population-level summary: Difference in means between treatment conditions.

Remarks:

(i) The "treatment policy strategy" for the intercurrent event "adherence to the investigational intervention" is consistent with an Intention-To-Treat principle/analysis discussed in Section 5.1. In our context, "adherence" is defined as "continue antibiotic treatment" as long as defined by the treatment strategy (not shorter, not longer).

(ii) It might happen that the blood culture (or other clinically relevant culture) turns positive late (after randomization), even though this is not expected to happen (or only in a few cases). If this happens, standard treatment that targets the newly identified cause of infection will be given identically in both arm. This is implicitly meant by the statement "regardless of any subsequent treatment decision or intervention during the follow-up needed to ensure good clinical care" in the treatment and intercurrent event attributes above.

4.2 Secondary endpoints

There are five secondary outcomes:

S1: Severe adverse event (SAE) within 28 days, defined as any of these events:

- Severe sepsis, defined as either (i) vasoactive treatment requirement or (ii) fluid bolus requirement >20mL/kg or (iii) respiratory support requirement (mechanical ventilation).

- Fever with positive blood culture. For possible contaminants, at least two positive culture bottles were required. Possible contaminants include coagulase negative staphylococcus, micrococcus spp., non hemolytic streptococcus, Corynebacterium spp, Proprionum spp, Bacillus spp.
- Severe focal infection highly suspicious of bacterial cause, requiring antibiotic treatment (e.g. lungs, abdomen, CNS, skin)
- Death

S2: New episode of neutropenic fever within 28 days.

S3: Days alive without fever within 28 days.

S4: Mortality within 28 days.

S5: Time to bone marrow recovery (up to 3 month), defined as the time from start of treatment to first absolute neutrophil count (ANC) $\geq 0.5 \times 10^9$ cells/L.

Note: see Section 5.7 for details about the definitions of fever, neutropenia/neutropenic, **S2** and **S3**.

4.3 Secondary objectives and estimands

For each of the secondary **binary** outcomes **S1**, **S2** and **S4**, the secondary objectives are to estimate the risks of the outcome in the two arms; as well as the risk differences. The corresponding clinical questions of interest are:

"What are the risks of [SAE / new episode of neutropenic fever / death] within 28 days after treatment started, when receiving the "short" and "long" treatment duration strategies, among children with cancer who started antibiotic treatment for high risk neutropenic fever of unknown origin, regardless of any intervention during the follow-up needed to ensure good clinical care? What is the corresponding risk [ratio / ratio / difference]?"

Accordingly, the corresponding estimands are defined as the primary estimands of Section 4.1, except for:

- the endpoint attributes, which are defined by the above items **S1**, **S2** and **S4**
- The intercurrent event "death within 28 days" does not exist for outcome **S4**. For **S1** and **S2**, it is implicitly addressed by the definition of the endpoint (as a classical competing event: in case death happens before the outcome of interest, the outcome of interest did not happen).
- Population-level summary: risk ratio for **S1** and **S2** and difference in risk for **S4** (between arms).

For each of the secondary **quantitative** outcomes **S3** and **S5**, the main secondary objectives are to estimate the medians of the outcome in the two arms; as well as their differences. The corresponding clinical questions of interest are:

"What are the median [number of day alive without fever without 28 days / time to bone marrow recovery], when receiving the "short" and "long" treatment duration strategies, among

children with cancer who started antibiotic treatment for high risk neutropenic fever of unknown origin, regardless of any intervention during the follow-up needed to ensure good clinical care? What is the corresponding risk difference?"

Accordingly, the corresponding estimands are defined as the primary estimands of Section 4.1, except for:

- the endpoint attributes, which are defined by the above items **S3** and **S5**
- with **S5**, the intercurrent event "death within 28 days" is implicitly addressed by the definition of the endpoint as a classical competing event: in case death happens before the outcome of interest, the outcome of interest did not happen and the time to bone marrow recovery is set to infinity to define the median time (see e.g. [20, 2]).
- Population-level summary: Difference in median between arms.

4.4 Exploratory outcomes

Exploratory outcomes are:

E1: Mortality within 3 months.

E2: New initiation of antibiotic treatment within 28 days, defined as antibiotic treatment restarted after at least 12 hours of stopping.

Remark: outcome **E2** can be thought as a composite outcome with "new episode of neutropenic fever" (**S2**) being one component, as a neutropenic fever systematically triggers initiation of antibiotics.

5 General Considerations

5.1 Intention-to-treat analyses

The main analyses supporting the primary and secondary objectives will use the "All episodes analysis set" that consists of all randomized episodes and correspond to Intention-To-Treat (ITT) analyses.

The ITT principle/analysis/estimand is often recommended to alleviate concerns about baseline confounding when analyzing the data. However, one drawback of the ITT estimand is that its magnitude depends on the degree to which participants receive the assigned interventions. See e.g., [6, 12]. In this trial, we expect that this is a minor issue, as good adherence is expected. Adherence data will be reported, see Section 10.2.3.

No Per Protocol analyses/estimands are pre-specified in this SAP, because non-adherence is expected to be negligible. However, post hoc analyses might be performed, e.g., if adherence data are not as good as anticipated. If performed, post hoc Per Protocol analyses are expected to be based on the "Per protocol analysis set" listed in Section 6.

5.2 Statistical Hypotheses

The null hypothesis to be tested in relation to the primary estimand (detailed in Sec. 4.1) is as follows:

- **Null hypothesis:** the mean number of days alive without antibiotic treatment within 28 days after treatment started is **the same** when the children receive the "short" and "long" treatment duration strategies[1]. Formally, $\mathcal{H}_0 : \mu_1 = \mu_0$, where μ_1 and μ_0 are the mean number of days when the children receive "short" and "long" treatment duration strategies, respectively.

versus

- **Alternative hypothesis:** the mean number of days alive without antibiotic treatment within 28 days after treatment started is **different** when the children receive the "short" and "long" treatment duration strategies[1]. Formally, $\mathcal{H}_1 : \mu_1 \neq \mu_0$.

We will use a two-sided test at 5% and matching 95% two-sided confidence interval (CI).

5.3 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. 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 [18].

5.4 Missing and interval censored data

No missing data are expected for the primary, secondary or exploratory outcomes. The patients are all children cancer patients that are monitored very closely, hence preventing missing outcome data. Missing data are not expected either for the baseline variables adjusted for in the primary analysis. In case missing data occur, this will be transparetnly reported as well as any change to the prespecified statistical analyses that they might cause.

Data on neutrophil counts, which are pivotal to define **S5** (time to bone marrow recovery), were not collected every day for all children. In accordance with the protocol, there were sometimes two or three days between consecutive blood samples, leading to interval censored data for some children. For instance, interval censoring [7, Sec. 16] can happen if a child is observed without bone marrow recovery a Friday (i.e., with neutrophil count $< 0.5 \times 10^9$ cells/L) and come back for the next blood sample only on Monday, when bone marrow recovery is observed (i.e., neutrophil count $\geq 0.5 \times 10^9$ cells/L). In that case, we only know that bone marrow recovery occurred between Friday and Monday. Hence, specific methods for interval censored data will be used to analyse outcome **S5** (see Section 8.2)

[1] among children with cancer who started antibiotic treatment for high risk neutropenic fever of unknown origin, regardless of any intervention during the follow-up needed to ensure good clinical care.

5.5 Within-patient correlation between episodes

A child can be randomized more than once and consequently a child can contribute with more than one episode. Bone marrow recovery and subsequent chemotherapy treatment must have taken place before a new randomization of the same child can occur, as detailed in the protocol. We cannot rule out that the outcome from two episodes from the same patient are more likely similar than the outcome of two episodes from two different patients. That is, we cannot rule out within-patient correlation between episodes. To analyze the data accordingly, we will use Generalized Estimating Equations (GEE) [17] as implemented in the popular `geepack` package of `R` [13]. Standard errors will be computed via either a standard "sandwich" approach, as implemented in `geepack`, or non-parametric cluster bootstrap [4, Sec. 3.8] (see Sections 7 and 8).

5.6 Covariate adjustment

All estimands considered in this SAP are "unconditional treatment effect" [10]. To follow usual guidelines and to attempt to gain power [10], we will adjust for a few baseline covariates in the analysis of the primary outcome. As a linear model without interaction will be used, no standardization step will be used, since "unconditional" and "conditional" effects are the same in that case [10]. As this would not be the case for the analysis of the secondary (binary) outcomes and also because of the limited sample size and limited prognostic value of the baseline variables for the secondary outcomes, the analysis of the secondary outcomes will not use covariate adjustment. Although parametric modeling assumptions will be made for covariate adjustment, randomization provides important robustness properties to type-I error control [26].

5.7 Pivotal definitions

5.7.1 Time zero

To define many outcomes, e.g., **S2**: "new episode of neutropenic fever within 28 days", we need to precisely define when the time starts. This is necessary to properly define the time window, e.g. "within 28 days". In this SAP, we define that time (i.e., time zero) at the time of treatment initiation, not at time of randomization, which should occur appropriately 48 hours later. This is consistent with the definition of the primary outcome (see Section 4.1) and it should facilitate the interpretation of the results.

5.7.2 Duration of treatment

Treatment durations are computed as the time difference between two dates and times: that of treatment stop and that of treatment start. Hence, treatment durations expressed in days are not whole numbers, but decimal numbers. Time of start and stop are defined as the first and last times when the treatment was administered to the patient. This definition of treatment duration is used to define the primary outcome and in the computation of some rates (See Section 10.1.1).

5.7.3 Fever and episode of fever

Fever is pivotal to define **S2** and **S3**. We say that a child has fever when the child fulfills (at least) one of these two conditions: a) a single temperature $\geq 38.5^{\circ}$ C once or b) temperature $\geq 38.0^{\circ}$ C sustained within one hour. An episode of fever starts as soon as we observed condition

a) or b). That is, as soon as a single temperature ≥ 38.5 is observed or once we have observed a temperature ≥ 38.0 for one hour (whichever comes first). An episode of fever stops 48 hours after the last time we observe the child with fever. That is, 48 hours after the last time a temperature ≥ 38.5 is observed or 48 hours after the last we have observed a temperature ≥ 38.0 for one hour, whichever comes last. Accordingly, a new episode of fever cannot start before 48 hours after the last time we observed condition a) or b). The temperature was monitored very regularly (≥ 1 per day) as is standard in this population of children with cancer.
To define **S3** (Days alive without fever within 28 days), we sum the duration of all episodes of fever within 28 days. A fever duration is computed as the time difference between two dates and times: that of fever stop[2] and that of treatment start. Hence, fever durations expressed in days are not whole numbers, but decimal numbers.

5.7.4 Neutropenia and episode of neutropenic fever

To define **S2** (New episode of neutropenic fever within 28 days), we also need to define neutropenia. Start of neutropenia is defined as the time when neutrophil counts are first observed $< 0.5 \times 10^9$. Time of stop of neutropenia is defined as the first time neutrophil count is observed $\geq 0.5 \times 10^9$ after start time. An episode of neutropenic fever starts when the child is first observed with both neutropenia and fever. It stops when either fever or neutropenia stops (whichever comes first).

Durations of neutropenia expressed in days are computed as the time difference between two dates and times: that of neutropenia stop and that of neutropenia start. Hence, treatment durations expressed in days are not whole numbers, but decimal numbers. This definition of neutropenia is used to compute secondary outcome **S2** and some rates (See Sections 4.2).

6 Analysis Sets

- The **"All episodes analysis set"** consists of all randomized episodes.
- The **"Per protocol analysis set"** consists of all randomized episodes except those for which the child did not receive the "short" or "long" treatment strategy to which he or she was randomized. These children will be identified as detailed in Section 10.2.3.

7 Analyses supporting the primary objective

The main analysis supporting the primary objective and corresponding to the estimand detailed in Section 4.1 will use the "All episodes analysis set" and corresponds to an ITT analysis.

To estimate the between arm difference in means with a 95% confidence interval and a p-value, we will use a multiple linear model (aka ANCOVA). The model will be fitted via GEE [17] and the standard errors will be computed accordingly, to account for possible within-patient correlation, as already mentioned in Section 5.5. We will use an independent working correlation structure, as we expect the correlation to be negligible and because additional assumptions might be needed for other choices [24]. The variance and link functions that define the GEE will be that of a gaussian family with an identity link, as commonly used for quantitative outcomes. Baseline covariates included in the model will be:

[2] pricisely, that of fever stop or 28 days after treatment initiation, whichever comes first (because we compute the sum within 28 days).

- Age at randomization
- Type of cancer ("AML or Relapse ALL", "ALL, in induction treatment", "ALL, not in induction treatment" or "Other cancers")
- Absolute neutrophil count (ANC) at presentation (≤ 0.1, $]0.1; 0, 5[$ or ≥ 0.5; unit is 10^9cells/L)
- Thrombocyte count at presentation
- Study site (Copenhagen, Aarhus or Odense)

These variables will be included in the model in addition to the binary variable that indicate the randomized assignment to "short" or "long" treatment duration. No interaction will be used in the multiple linear model. To avoid unnecessarily strong linearity assumptions, the effects of age and thrombocyte count will be modelled via linear splines, with two knots at 6 and 13 years for age and two knots at 50 $\times 10^9$cells/L and 100 $\times 10^9$cells/L for thrombocyte count. An example of `R` code corresponding to this analysis is provided in appendix A.1.

8 Analyses supporting secondary objectives

The main analysis supporting the secondary objectives and corresponding to the estimands detailed in Section 4.3 will use the "All episodes analysis set" and corresponds to an ITT analysis [16].

8.1 Binary secondary endpoints

Secondary outcomes **S1**, **S2** and **S4** are binary. For **S1** and **S2**, we will first compute the logarithm of the risk ratio and its confidence interval using a GEE approach [17], again to account for a potential within-patient correlation. Unlike for the primary analysis, here we will use a log-linear model. Specifically, the variance and link functions that define the GEE will be that of a binary family with a log link. Here again, we will use an independent working correlation structure, for the same reasons mentioned above. To compute the point estimate of the risk ratio and its confidence interval, we will then exponentiate the fitted log of the risk ratio and its confidence interval. We will not adjust for baseline variables in this analysis, mainly for three reasons already mentioned in Section 5.6. First, non-linear models do not directly provide marginal effect estimates [10]. Second, the baseline covariates are not expected to be substantially prognostic of the outcomes and third, because of the limited (effective) sample size. An example of `R` code corresponding to this analysis is provided in appendix A.2.

For outcome **S4** (mortality), we expect very few events (if any), say less than 1% (so, ≤ 2 in the dataset). Hence we will use an exact method instead of a GEE approach, as the asymptotic approximations implicitly used by a GEE approach would be very questionable. The method will not account for the within patient correlation, as we are not aware of any exact method that could do that. But, this is a relatively minor limitation when the risks are very small. Specifically, we will use an exact unconditional test using the score statistic ordering as in Nielsen et al. [19], and implemented in the `uncondExact2x2()` function of the `R` package `exact2x2`; see [8] for the mathematical details and Appendix A.3 for the specific `R` code that we will use.

8.2 Quantitative secondary endpoints

Secondary outcomes **S3** (days alive without fever within 28 days) and **S5** (time to bone marrow recovery) are quantitative outcomes. They will be analyzed by estimating the median within each arm and their difference, together with two-sided 95%-CIs. The standard errors to compute the CIs will be computed using non-parametric cluster bootstrap (using B=1000 replicates). This means that we will bootstrap the patients, not the episodes, following "Strategy 1" in Section 3.8 of Davison and Hinkley [4]. An example of `R` code corresponding to this analysis for secondary outcome **S3** is provided in appendix A.4.

Unlike for **S3**, for secondary outcome **S5** (time to bone marrow recovery) the estimator of the median that we will use will not be the standard one. That is, we will not use the `median()` function of `R` as described in Appendix A.4. This is because we expect to observe interval censored data, as already discussed in Section 5.4. Instead, we will proceed as follows. First, we will use an appropriate and well known method to estimate the cumulative incidence function from interval-censored data: the nonparametric maximum likelihood estimator method implemented in the `icfit()` function of the `R` package `interval` (see [9] or [7, Sec. 16.5] and references therein). Second, from the estimated cumulative incidence function, we will then estimate the median as $\inf\left\{t : \widehat{F}(t) \geq 0.5\right\}$, where $\widehat{F}(t)$ denotes the estimated cumulative incidence function at time t [2].

9 Analyses supporting exploratory outcomes

The main analyses of the exploratory outcomes detailed in Section 4.4 will also use the "All episodes analysis set" and corresponds to ITT analyses. They will be analyzed using the same methods as those used for the secondary outcomes, as outlined in Section 8.1. Specifically, **E1** (mortality within 3 months) will be analyzed as secondary outcomes **S4** (mortality within 28 days) while **E2** (new initiation of antibiotic treatment within 28 days) will be analyzed as secondary outcomes **S2** (new episode of neutropenic fever within 28 days).

10 Other analyses

10.1 Additional outcome analyses

10.1.1 Rates of SAE during neutropenia

We will compute the rate of SAE during neutropenia, within 28 days from treatment initiation, among children that are:

(i) currently treated with antibiotics

(ii) not currently treated with antibiotics.

(iii) not currently treated with antibiotics, but have stopped antibiotic treatment less than three days ago.

For the primary analysis of these rates, we will pool the data of the two arms. The rationale for pooling the data is that antibiotic treatment are expected to have a short term effect. Hence, it seems reasonable to assume that a child currently treated with antibiotics is approximately similarly protected no matter how long ago the treatment has started, either many days ago (as typical in

the "long" arm) or just a few days ago (as typical in the "short" arm). We will compare the rates via rate ratios. The rate ratio of main interest is that comparing the rates of the first and third groups. Indeed, an interesting research question is: *"Is there an increased risk of SAE just after stopping the antibiotic treatment during neutropenia?"* (here "just after stopping" is defined as "within three days from stopping"). Comparing the rates of the first and second groups is also of interest.

To compute the rates, we will divide the number of events by the total time duration (days) at risk, as commonly done. Note that by definition of the rate of interests (during neutropenia), a time period (day) is not "at risk" if ANC is not $< 0.5 \times 10^9$cells/L. Time at risk will be expressed in days, but computed as detailed in Section 5.7.2. This means that a continuous time scale will be used, not a discrete one. Exact Poisson inference will be used to compute 95% CI for the rates and the rate ratios using the `poisson.test()` function of `R` . Specifically, we will look at the rate of the first SAE. Although SAEs is a potentially recurrent event, we will investigate the rate of the first of possibly several SAEs. Descriptive statistics about children experiencing more than one SAE (if any) will be provided for completeness. We will consider presenting a descriptive plot similar to that of Figure 1 to transparently report the raw data about numbers of events and days at risk and show how these analyses summarize these raw data.

For completeness, we will also redo the analyses stratified per arm (i.e., without pooling the data).

10.1.2 Rates of new episode of fever during neutropenia

We will also perform an analysis of the rate of a new episode of fever during neutropenia, among afebrile children. We will compare the rates between the same three treatment group as for the analysis of the rates of SAE presented in Section 10.1.1, using the same methodology.

10.1.3 Cumulative incidence plots

To describe the timing of the binary secondary outcome **S1** and **S2**, we will report the cumulative incidence/risk function per arm graphically and with pointwise 95% CIs. That is, we will present the proportion of events observed before day t, for $t = 0, 1, 3, \ldots, 28$, for each arm. The CIs will be computed again using GEE. A similar approach will be used for **S4**, if relevant (i.e., if several deaths are observed).

Cumulative incidence functions per arm will also be reported for the time to bone marrow recovery (**S5**). However, the method to compute them will be that accounting for interval censoring outlined in Section 8.2.

10.1.4 Subgroup analyses

There is only one pre-specified subgroup analysis.

- episodes with ANC at presentation $\leq 0.1 \times 10^9$ cells/L, i.e., episodes of patients considered at "very high" risk.

We will replicate the main analysis of the primary outcome described in Section 7, up to minor differences for the baseline covariate adjustment. Specifically, for any baseline categorical variable, a group of less than five episodes will be combined with the smallest group of more than five episodes. Similarly, we will also drop knots if the subgroups that they implicitly define do not

all contain at least than five episodes (0-5, 6-12 and $\geq$13 years for age and 0-49, 50-99, $\geq$ 100 for thrombocyte, in 10^9cells/L unit). We will also replicate the analysis of main analyses of the secondary outcomes described in Section 8.

10.2 Descriptive analyses

10.2.1 Recruitment

Recruitment of the patients and episodes 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]. Number of patients, number of episodes and the distribution of number of episodes per patient will be provided.

10.2.2 Baseline characteristics

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

- Sex (male/female)
- Age (year)
- Study site
- Year of inclusion (Nov 2020 to Dec 2021, 2022, 2023 or 2024)
- Cancer disease and treatment cycle (ALL induction, ALL not induction, AML, Relapsing ALL, Lymphoma, Neuroblastoma, Other)
- Days from cancer diagnosis to randomization
- Empirical antibiotic treatment at randomization
- ANC at presentation
- Minimum ANC within the time interval ranging from presentation to randomization
- Thrombocyte count at presentation
- Neutropenia duration at randomization

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].

10.2.3 Adherence

For each arm, we will document the number of patients who did not receive the "short" or "long" treatment strategy to which they were randomized. Specifically, we will report the number and proportion of:

- patients randomized to "short" who continued antibiotics longer than they should according to the "short" duration treatment strategy.
- patients randomized to "long" who stopped antibiotics before they should according to the "long" duration treatment strategy.

The reasons for that will be provided too, as these data have been collected. Note that the relevant data have been collected to identify such patients. Especially, information about the time at which the treatment should have stopped has been collected. That is, the time corresponding to:

- 48h of apyrexia and stability (for patients in the "short" arm)
- either i) neutrophil count is $\geq 0.5 \times 10^9$ cells/L, and the child is afebrile and clinical stable or ii) the child has received 10 days of antibiotics and have been afebrile and clinically stable for 7 days (for patients in the "long" arm).

Daily data about antibiotic treatment have been collected too. Hence, it is possible to check that the random assignment to "long" or "short" is consistent with the daily data about antibiotics and the time at which the treatment should be discontinued. Any inconsistency of more than 24h will be reported, but any data "inconsistency" of less than 24h will not be considered as an inconsistency. This is because such small inconsistencies are expected because of logistical constraints. For instance, small inconsistencies might happen for patients treated with out-patient intravenous antibiotics by portable infusion pumps according to local guidelines. These patients need to come back to the hospital regularly for filling of pumps. For example, a patient needs to come back to the hospital to have the pump that administers antibiotics removed (hence continuing the treatment slightly longer than planned). A patient expected to fulfill the criteria to stop antibiotics within less than 24h might also not receive a new pump (hence stopping slightly earlier than planned).

10.2.4 Type of severe adverse events

To complement the analysis of **S1** (any severe adverse event), we will present the number and proportion of each individual component of the composite outcome **S1**, per arm. The individual component are listed in Section 4.2 to define **S1**. The type of observed severe focal infections will also be listed (e.g., lungs, abdomen, skin).

10.2.5 Descriptive statistics of outcomes

Outcomes (primary, secondary and exploratory) will also be descriptively summarized, per arm, using the "All episodes analysis set". 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. The number and proportions of missing data will be reported if any, but we do not expect any.

10.2.6 Dotplots and boxplots of the primary outcome

The raw data of the primary outcome will be presented graphically via dotplots (one per arm). Boxplots will be overlaid, to summarize the raw data in each group graphically.

10.2.7 Descriptive plot of antibiotics, neutropenia and more

For each of the 28 first days of each patients, a plot will show the days:

- with and without antibiotics
- with or without neutrophil count $< 0.5 \times 10^9$ cells/L

An example of such a plot is Figure 1. We will consider enriching the plot by additionally showing the days with and without fever and the occurrences of SAEs. The main results of the trial will essentially be numerical summaries of this exhaustive graphical representation, which should provide the "full picture".

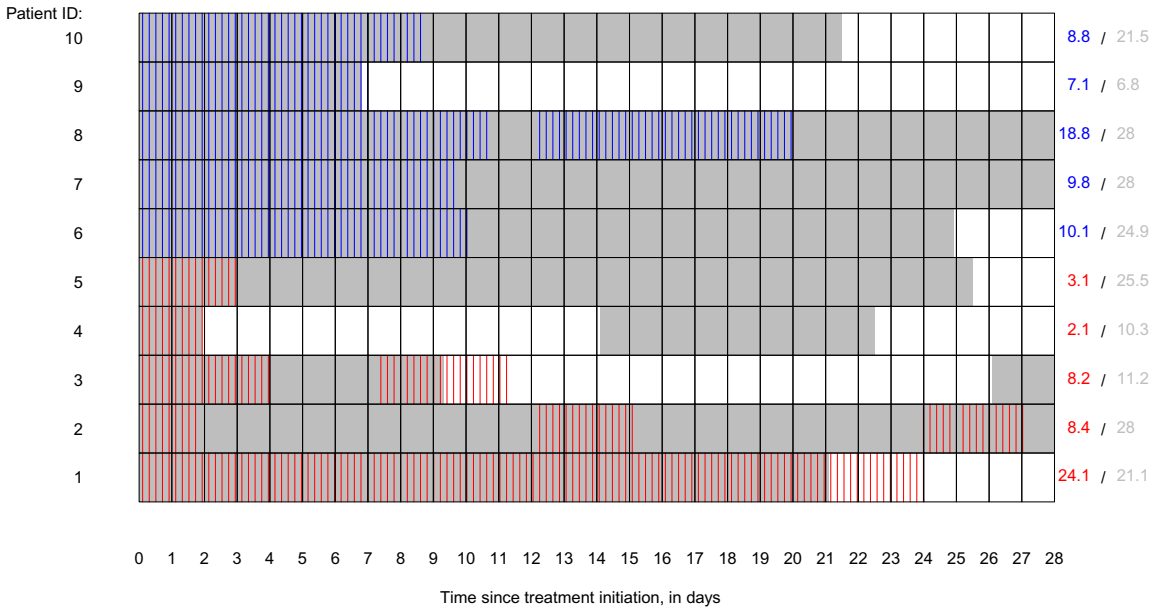


Figure 1: Example of descriptive plot to show the days with antibiotics (in blue when randomized to "long" duration; red when randomized to "short") and neutropenia (grey) for each patient, within the 28 days follow-up. The plot shows the data for 5 hypothetical patients in each arm. Total number of days with antibiotics per patients are shown in the right margin (in color), together with the number of days with neutropenia (in grey). Note that treatment and neutropenia can start and stop at any time of the day and these times have been recorded (as shown in the plot). Hence the durations in days in the right margin are not whole numbers.

10.2.8 Regularity of neutrophil count measurements

Secondary outcome **S5** (Time to bone marrow recovery) is interval censored as it is observed from repeated measures of neutrophil counts (see Section 5.4). Descriptive statistics about the regularity of the data collection will be presented per arm, to document how interval censoring impacts the analysis of the secondary outcome **S5**.

10.2.9 Blood culture turning positive after randomization

Inclusion criteria include the fact that the blood culture (or other clinically relevant culture) was not positive after 48h (i.e., "negative"). Very few culture are expected to become positive after 48h ($< 5\%$). Hence, we expect that the culture of very few randomized episodes will become positive. For completness, we will report the number and proportion per arm. Descriptive statistics about the time at which the culture becoming positive will also be described.

11 Sample size determination and power calculations

11.1 Sample size determination

For the number of episodes to include in the trial, the initial sample size calculation was performed as followed. A mean number of 16 empirical antibiotic therapy free days was expected in the experimental group, versus 13 days in the control group. A standard deviation of 6.5 days in each arm was expected. These expected values were, among other things, based on previous results from a similar study in adult patients [1]. Standard sample size calculation for 1:1 randomization [3, Sec. 3.2.1] suggested that

$$n/2 = \frac{2 \times 6.5^2 \times (1.96 + 1.282)^2}{(16 - 13)^2} \approx 99$$

in each group, hence $n = 198$ in total, would be sufficient to obtain 90% power to show a significant difference in mean between the two arms, while controlling the type-I error at 5% (using a usual two-sided test). Accounting for up to 10% of drop-outs, the sample size was increased to $n = 198/(1 - 0.1) = 220$.

11.2 Additional power calculations

After the decision to stop the trial early was taken, it has been anticipated that about 100 episodes would be included by the end of the trial. Below we provide the results of a standard power calculation [3, Sec. 3.2.1] for a sample size (i.e. number of episode, n) ranging from 90 to 110 (and n=198), assuming the same difference in means and standard deviation as for the initial sample size calculation detailed in Section 11.1 (and no dropout, as none were observed and dropouts were no longer expected).

Sample size (n)	90	95	100	105	110	198
Power	59%	62%	64%	65%	68%	90%

Further calculation suggest that, to get 90% power with a sample size of n=100 while assuming a standard deviation of 6.5 days, we would need to assume a difference in means of 4.3 days instead 3 days. Assuming a more modest difference in means of 3.5 days, the power is computed as 77%.

12 Changes to protocol-planned Analyses

1. Time to next chemotherapy is no longer a secondary outcome. It is no longer considered a secondary outcome of interest for two main reasons. First, it is influenced by numerous factors beyond infections, such as logistical and scheduling issues which introduces variability. Second, this outcome is essentially a proxy for the more relevant "Time to bone marrow recovery", which already serves as a secondary outcome (**S5**) and more directly reflects patient recovery and treatment effectiveness.

2. Alterations in gut microbiome is no longer considered a secondary outcome. It is now considered as an exploratory outcome that will be analyzed later, hence this SAP does not pre-specify its analysis.

3. The initial statistical analysis plan mentioned a non-inferiority margin of 10% for the secondary outcomes analysis. This was not the results of carefully considerations and we will no longer consider this margin. It is irrelevant to non-binary secondary outcomes (**S3** and **S5**) and its clinical relevance in unclear for the analysis of binary outcomes (**S1**, **S2** and **S4**). Consequently, for the interpretation of the results, there will be no specific focus on whether the two-sided CIs contain a specific non-inferiority margin.

4. The initial statistical analysis plan mentioned a different set of baseline covariates to adjust for in the analysis of the primary outcome. This was not the results of very careful considerations, but first thoughts to mimic what has been done in a similar study [1]. New careful thinking and discussions led to the conclusion that the covariates listed in Section 7 are more relevant. Of note, adjusting on thrombocyte was thought relevant based on, among other things, results from [15]. Adjusting for site was also thought relevant, as randomization was stratified by site. The decision to update the list of covariates was not based on any preliminary analysis of the outcome, as usually recommended [5].

5. Exploratory outcomes (**E1** and **E2**) and other analyses, such as the rate analyses (Section 10.1.1), have been added.

References

[1] Aguilar-Guisado, M., Espigado, I., Martín-Peña, A., Gudiol, C., Royo-Cebrecos, C., Falantes, J., Vázquez-López, L., Montero, M. I., Rosso-Fernández, C., de la Luz Martino, M., et al. (2017). Optimisation of empirical antimicrobial therapy in patients with haematological malignancies and febrile neutropenia (how long study): an open-label, randomised, controlled phase 4 trial. *The Lancet Haematology*, 4(12):e573–e583.

[2] Beyersmann, J., Friede, T., and Schmoor, C. (2022). Design aspects of covid-19 treatment trials: improving probability and time of favorable events. *Biometrical Journal*, 64(3):440–460.

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

[4] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.

[5] EMA/CHMP (2015). Guideline on adjustment for baseline covariates in clinical trials. Technical report, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf.

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

[7] Fay, M. P. and Brittain, E. H. (2022). *Statistical Hypothesis Testing in Context*. Cambridge University Press.

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

[9] Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. *Journal of Statistical Software*, 36(2).

[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] 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.

[12] Hernán, M. A. and Scharfstein, D. (2018). Cautions as regulators move to end exclusive reliance on intention to treat.

[13] Højsgaard, S., Halekoh, U., and Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of statistical software*, 15:1–11.

[14] 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.

[15] Jackson, T. J., Napper, R., Haeusler, G. M., Pizer, B., Bate, J., Grundy, R. G., Samarasinghe, S., Angelini, P., Ball-Gamble, A., Phillips, B., et al. (2023). Can i go home now? the safety and efficacy of a new uk paediatric febrile neutropenia protocol for risk-stratified early discharge on oral antibiotics. *Archives of Disease in Childhood*, 108(3):192–197.

[16] Kahan, B. C., White, I. R., Edwards, M., and Harhay, M. O. (2023). Using modified intention-to-treat as a principal stratum estimator for failure to initiate treatment. *Clinical Trials*, 20(3):269–275.

[17] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[18] 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.

[19] Nielsen, A. B., Holm, M., Lindhard, M. S., Glenthøj, J. P., Borch, L., Hartling, U., Schmidt, L. S., Rytter, M. J., Rasmussen, A. H., Damkjær, M., et al. (2024). Oral versus intravenous

empirical antibiotics in children and adolescents with uncomplicated bone and joint infections: a nationwide, randomised, controlled, non-inferiority trial in denmark. *The Lancet Child & Adolescent Health*, 8(9):625–635.

[20] Peng, L. and Fine, J. (2007). Nonparametric quantile inference with competing–risks data. *Biometrika*, 94(3):735–744.

[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] Sullivan Pepe, M. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in statistics-simulation and computation*, 23(4):939–951.

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

[26] Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M., and Rosenblum, M. (2023). Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, 118(542):1152–1163.

A Appendix

A.1 R code for main analysis

```
# create variables to fit with linear splines
d$age5 <- as.numeric(d$age>5)*(d$age-5)
d$age12 <- as.numeric(d$age>12)*(d$age-12)
d$thrombo50 <- as.numeric(d$thrombo>50)*(d$thrombo-50)
d$thrombo100 <- as.numeric(d$thrombo>100)*(d$thrombo-100)
# fit model
library(geepack)
geePrimary <- geeglm(days~ treat + age + age5 + age12
                     + cancer + anc + thrombo + thrombo50 +
                     thrombo100 + site,
                     data=d,                      # data set (long format)
                     id=id,                       # patient ID (clusters)
                     family=gaussian("identity"), # usual linear model parametrization
                     corstr="independence")       # independent working correlation structure
summary(geePrimary)
```

In the above, the dataset "d" is assumed to include the following variables: "days" indicates the number of days alive without antibiotic treatment within the first 28 days, "id" the patient ID (clustering variable), "treat" the randomized treatment strategy, "age" the age at randomization, "cancer" the type of cancer, "anc" the absolute neutrophil count at presentation, "thrombo" the thrombocyte count at presentation and "site" the study site.

A.2 R code for secondary binary outcomes S1 and S2 analysis

```
library(geepack)
geeBinary <- geeglm(Y~treat,
                    data=d,                          # data set (long format)
                    id=id,                           # patient ID (clusters)
                    family=binomial("log"),          # log-linear model for a binary outcome
                    corstr="independence")           # independent working correlation structure
#--- extract estimate and SE (log-scale) ------------
logRR <- summary(geeBinary)$coef["treat","Estimate"]
SElogRR <- summary(geeBinary)$coef["treat","Std.err"]
#--- compute Risk Ratio and 95% CI ------------------
Results <- exp(c(logRR,
                 logRR-qnorm(0.975)*SElogRR,
                 logRR+qnorm(0.975)*SElogRR))
```

In the above, the variable "Y" indicates the binary outcome of interest, "id" the patient ID (clustering variable) and "treat" the randomized treatment strategy.

A.3 R code for mortality analysis (S4)

```
library(exact2x2)
uncondExact2x2(x1=x1obs,                  # x1obs is the number of deaths in the "short" arm
               n1=n1obs,                  # n1obs is the number of episodes in the "short" arm
               x2=x2obs,                  # x2obs is the number of deaths in the "long" arm
               n2=n2obs,                  # n2obs is the number of episodes in the "long" arm
               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
```

A.4 R code for secondary quantitative outcome S3

In the R code below, we assume that the dataset R contains the following variables:

- days: the outcome, a number of days corresponding to either S3 or S5.
- arm: the randomized arm ("short" or "long", coded as 0/1)
- Patient: the patient ID, to link all the episodes from the same patient.

```
# First we compute the median in each arm
med1 <- median(d$days[d$arm==1])
med0 <- median(d$days[d$arm==0])
# Then we bootstrap to compute SE and 95%-CI;
# we set the seed for reproducibility.
set.seed(20240930)
```

```
ResB <- matrix(NA,ncol=2,nrow=1000)
AllPatientsID <- unique(d$Patient)
for(b in 1:nrow(ResB)){
    # Sample the patients with replacement
    idb <- sample(x=AllPatientsID,
                  size=length(AllPatientsID),
                  replace=TRUE)
    # Create a corresponding bootstrapped data set, by stacking
    # the data of all episodes of each (possibly duplicated)
    # patient that has been  sampled.
    db <- NULL
    for(theid in idb){
        db <- rbind.data.frame(db,d[which(d$Patient==theid),])
    }
    # Compute and save medians per arm from each bootstrapped data
    ResB[b,1] <- median(db$days[db$arm==1])
    ResB[b,2] <- median(db$days[db$arm==0])
}
# Compute the standard error for the difference in medians
SE <- sd(ResB[,1]-ResB[,2])
# Compute the corresponding 95% confidence interval (Wald-type)
results <- c(Est=med1-med0,
             lower=med1-med0-qnorm(0.975)*SE,
             upper=med1-med0+qnorm(0.975)*SE)
```