Equivalence trials for sepsis in newborns and young infants

Review of the statistical aspects of trials claiming to show equivalence of simplified treatment regimes
**ABSTRACT**

Serious bacterial infections are a major cause of neonatal and young-infant deaths worldwide. Identification is difficult, and standard-of-care treatment is demanding. Several trials have examined simplified treatment regimes that include oral antibiotics. In this report, four recently published randomized trials were analysed for their statistical power to demonstrate equivalence of simplified antibiotic regimes with the standard of care among children with bacterial infection. These trials were found to be underpowered in terms of both overall sample size and number of included children with bacterial infection. They cannot support the conclusion that the alternative antibiotic regimes studied are equally effective. Much larger sample sizes and more selective inclusion criteria are needed. The conclusion that simpler treatment regimens for the treatment of newborn sepsis are effective is not warranted, based on the published trials. The promotion of such regimens as an alternative to established treatments needs to be stopped.

**Keywords**

NEWBORN  
NEONATAL SEPSIS  
INFANT  
SIGNS AND SYMPTOMS
Equivalence trials for sepsis in newborns and young infants

Review of the statistical aspects of trials claiming to show equivalence of simplified treatment regimes
Acknowledgements ................................................................. vi
Abbreviations ................................................................. vi
Introduction ................................................................. 1
Methods ................................................................. 3
  Trials ................................................................. 3
  Inclusion/exclusion criteria and case-finding ................................ 7
  Prevalence of bacterial infection in infants eligible for recruitment ................................ 9
  Sensitivity and specificity of recruitment criteria ................................ 11
  Apparent treatment-failure rate among infants without bacterial infections ................................ 11
  Statistical methods ................................................................. 11
Results ................................................................. 13
  Corrected type I error rate, equivalence margin, and prevalence estimates ................................ 13
  Sample size and recruitment criteria accuracy needed to show equivalence in neonates and young infants with bacterial infections ................................ 13
  Visualization for power and sample size correction ................................ 14
  Interactive tools for computing corrected measures ................................ 16
Discussion ................................................................. 17
  Limitations ................................................................. 18
  Conclusions ................................................................. 19
References ................................................................. 20
Annex 1. Details of statistical calculations ................................ 22
Acknowledgements

This paper was written by Dr Emma Thomas, RAND Corporation, Santa Monica, United States of America, and Dr Pia Maier, Centre for Paediatric and Adolescent Medicine, University Hospital Heidelberg, Germany.

Statistical review of the paper was performed by Dr Denis Agniel, RAND Corporation, Santa Monica, United States of America.

The WHO Regional Office for Europe wishes to acknowledge the helpful discussions and critical review of the paper from: Professor John Carlin, Murdoch Children’s Research Institute and University of Melbourne, Australia; Professor Ron Dagan, University of Beer Sheva, Israel; Professor Trevor Duke, University of Melbourne, Australia; and Professor Kim Mulholland, London School of Hygiene and Tropical Medicine, United Kingdom.

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFRINEST</td>
<td>African Neonatal Sepsis Trial</td>
</tr>
<tr>
<td>IMCI</td>
<td>integrated management of childhood illness</td>
</tr>
<tr>
<td>MLE</td>
<td>maximum likelihood estimate</td>
</tr>
<tr>
<td>PPV</td>
<td>positive predictive value</td>
</tr>
<tr>
<td>pSBI</td>
<td>possible serious bacterial infection</td>
</tr>
<tr>
<td>SATT</td>
<td>Simplified Antibiotic Therapy Trial</td>
</tr>
</tbody>
</table>
Introduction

Serious bacterial infections are a major cause of neonatal deaths worldwide, including in the higher-mortality countries of the WHO European Region (1). Manifestations such as pneumonia, meningitis and sepsis form a continuum that is difficult to distinguish in this age group, so all are summarized as neonatal sepsis (2).

Over the past 30 years, WHO has invested in several studies on the aetiology of neonatal sepsis and clinical signs that might possibly identify neonates and young infants with sepsis (3–5). The first compilation of such signs derived from the WHO Young Infants Study undertaken in the early 1990s (3,4). The list comprised 14 signs, and this fact led to a high sensitivity but low specificity of the algorithm (2). This list of signs was integrated into the neonatal integrated management of childhood illness (IMCI) algorithm as signs indicating “possible serious bacterial infection” (6).

Because of the large number of children being referred following this approach, and both parents and health workers being reluctant to refer children who had one sign but were not considered sick, another large multicentre study was undertaken (5). The analysis of this study showed that the list of predictors of possible sepsis could be reduced to seven signs, maintaining a reasonable balance of sensitivity and specificity (sensitivity of 85% and specificity of 75% in this second study for neonates in the first week of life for indicating need for admission to hospital).

Both studies (2,5) were performed in ill infants who had been taken to a health facility. While the first study was interested mainly in predicting sepsis, the second was expanded to also identify infants with other conditions, such as birth asphyxia and jaundice. Even though this was a larger list of conditions, the term “possible serious bacterial infections” was maintained. Looking at the yield of blood cultures in these studies, only around 10% of the preselected children had a positive culture, many of which had possible contaminants (4,7).

Signs of bacterial infection are often non-specific and subtle, and neonates might deteriorate quickly. Even where a qualified assessment is possible, the standard of care is considered to be admission to hospital for observation and initiation of antimicrobial treatment while waiting for test results, then monitoring the child’s progress in response to the initiated treatment. Inpatient treatment usually consists of injectable antibiotics such as ampicillin and gentamicin. Oral treatment is not considered an option in this age group in developed countries (8).

Timely access to a competent health-care provider is often a problem where health systems are weak. Hospital treatment often is either not possible or not acceptable for families in resource-poor settings. They might not be convinced that the child is seriously ill, or the referral would put an undue financial or other burden on them. In addition, the poor reputation of many hospitals means they may not be seen as offering a good option for improving survival.

Without further work on clinical signs or diagnostic tests to identify children who really have a bacterial infection, a series of randomized controlled equivalence trials began to compare the standard of care with simplified antibiotic regimens administered in the community for young infants. These trials, called the African Neonatal Sepsis Trial (AFRINEST) and the Simplified Antibiotic Therapy Trial (SATT) (9–11), have been published within the last six years. They led to treatment recommendations for neonates and young infants in situations
“where referral is not possible” (12). The recommendations are often suggested as an alternative option for treatment of neonatal sepsis beyond established practice.

Some concerns have emerged about these trials. Equivalence trials of antibiotic treatments that include large numbers of neonates and young infants without bacterial infection are at high risk of erroneously concluding equivalence. This is the case even when there is a clinically meaningful difference between regimens for neonates and young infants with bacterial infection. This statistical liability derives from the fact that following randomization, all trial arms will have similar and large proportions of neonates and young infants without bacterial infection. These infants will not respond to antibiotic treatment of any kind, causing the spontaneous cure and failure rates to be roughly the same for all trial arms. As a result, the observed treatment difference between trial arms will be much smaller than the true treatment difference for infants with bacterial infections, who are the only infants who stand to benefit from antibiotic treatment. These observations are not new: other authors have discussed this effect and its consequences in detail for different bacterial infections, such as acute otitis media (13–18). The AFRINEST and SATT trials are particularly susceptible to this flaw since, as discussed, the inclusion criteria were based on clinical signs that have much less than perfect specificity for true bacterial infection.

This report shows that these trials of sepsis treatments in neonates and young infants were not able to demonstrate equivalence of antibiotic regimens as claimed. By adjusting key trial parameters to account for the likely proportion of randomized infants with true bacterial infections, it demonstrates that these trials were drastically underpowered to detect treatment equivalence. As a consequence, the true uncertainty in the estimated treatment difference is much higher than reported.

The report also discusses how the trial characteristics make the conclusion of treatment equivalence even less applicable to the general population of newborns and young infants for whom the recommendations “where referral is not possible” are made, even in high-mortality settings.
Methods

Trials

This section presents a summary of four randomized equivalence trials used for the development of the guidelines for antibiotic treatment of neonates and young infants when referral is not possible (12). These trials, summarized below, aimed to compare antibiotic regimens for the treatment of clinical severe infection in outpatient settings.

1. **The SATT trial in Pakistan**, which compared three treatment regimens (9):
   a. intramuscular procaine benzylpenicillin and gentamicin once per day for seven days;
   b. oral amoxicillin twice per day and intramuscular gentamicin once per day for seven days; and
   c. intramuscular procaine benzylpenicillin and intramuscular gentamicin both once per day for two days followed by oral amoxicillin twice per day for five days.

2. **The AFRINEST trial at five sites in the Democratic Republic of the Congo, Kenya and Nigeria**, which compared four treatment regimens (11):
   a. intramuscular procaine benzylpenicillin and gentamicin once per day for seven days;
   b. intramuscular gentamicin once per day and oral amoxicillin twice per day for seven days;
   c. intramuscular procaine benzylpenicillin once per day for two days followed by oral amoxicillin twice per day for five days; and
   d. intramuscular gentamicin once per day for two days followed by oral amoxicillin twice per day for seven days.

This part of the AFRINEST study included children with signs of infection other than fast breathing only.

3. **The AFRINEST trial at five sites in the Democratic Republic of the Congo, Kenya and Nigeria**, which compared two treatment regimens (10):
   a. intramuscular procaine benzylpenicillin and gentamicin once per day for seven days; and
   b. oral amoxicillin twice per day for seven days.

This part of the AFRINEST study included children with fast breathing only.

4. **An equivalence trial in Bangladesh**, which compared three treatment regimens (19):
   a. intramuscular procaine benzylpenicillin and gentamicin once per day for seven days;
   b. intramuscular gentamicin once per day and oral amoxicillin twice per day for seven days; and
   c. intramuscular procaine benzylpenicillin and gentamicin once per day for two days followed by oral amoxicillin twice per day for five days.

Features and results of the trials are summarized in Fig. 1.
Fig. 1. Summary of equivalence trials on neonatal sepsis

STUDY 1

- **A** Procaine benzylpenicillin and gentamicin (7 days) 747 infants
  - Deaths within 7 days 11 = 1%
  - Deaths before day 15 13 = 2%
  - Deaths by day 15 12 = 2%
  - Deaths before day 15 13 = 2%

- **B** Amoxicillin and gentamicin (7 days) 751 infants
  - Deaths within 7 days 7 = 1%
  - Deaths before day 15 9 = 1%
  - Deaths by day 15 7 = 1%
  - Deaths before day 15 9 = 1%

- **C** Procaine benzylpenicillin and gentamicin (2 days), then amoxicillin (5 days) 753 infants
  - Deaths within 7 days 10 = 1%
  - Deaths before day 15 12 = 2%
  - Deaths by day 15 10 = 1%
  - Deaths before day 15 12 = 2%

Source: Mir et al. (9).

STUDY 2

- **A** Procaine benzylpenicillin and gentamicin (7 days) 828 infants
  - Deaths by day 8 10 = 1%
  - Deaths by day 15 12 = 1%
  - Deaths by day 8 10 = 1%

- **B** Amoxicillin and gentamicin (7 days) 826 infants
  - Deaths by day 8 8 = 1%
  - Deaths by day 15 10 = 1%
  - Deaths by day 8 10 = 1%

- **C** Procaine benzylpenicillin and gentamicin (2 days), then amoxicillin (5 days) 862 infants
  - Deaths by day 8 17 = 2%
  - Deaths by day 15 20 = 2%
  - Deaths by day 8 17 = 2%

- **D** Gentamicin and amoxicillin (2 days), then amoxicillin (5 days) 848 infants
  - Deaths by day 8 46 = 5%
  - Deaths by day 15 11 = 1%
  - Deaths by day 8 46 = 5%

Source: Tshefu et al. (11).
Fig. 1 contd

### STUDY 3

<table>
<thead>
<tr>
<th>Group</th>
<th>Intervention</th>
<th>Duration (days)</th>
<th>N</th>
<th>PTF* by day 8</th>
<th>Deaths by day 8</th>
<th>Deaths within 15 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Procaine benzylpenicillin and gentamicin</td>
<td>7</td>
<td>1061 infants</td>
<td>234 = 22%</td>
<td>4 = &lt;1%</td>
<td>4 = &lt;1%</td>
</tr>
<tr>
<td>B</td>
<td>Amoxicillin</td>
<td>7</td>
<td>1135 infants</td>
<td>221 = 19%</td>
<td>2 = &lt;1%</td>
<td>4 = &lt;1%</td>
</tr>
</tbody>
</table>

Source: Tshefu et al. (10).

### STUDY 4

<table>
<thead>
<tr>
<th>Group</th>
<th>Intervention</th>
<th>Duration (days)</th>
<th>N</th>
<th>PTF* by day 8</th>
<th>Deaths before day 8</th>
<th>Deaths before day 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Procaine benzylpenicillin and gentamicin</td>
<td>7</td>
<td>795 infants</td>
<td>78 = 10%</td>
<td>13 = 2%</td>
<td>14 = 2%</td>
</tr>
<tr>
<td>B</td>
<td>Gentamicin and amoxicillin</td>
<td>7</td>
<td>782 infants</td>
<td>65 = 8%</td>
<td>9 = 1%</td>
<td>12 = 2%</td>
</tr>
<tr>
<td>C</td>
<td>Procaine benzylpenicillin and gentamicin (2 days), then amoxicillin (5 days)</td>
<td></td>
<td>790 infants</td>
<td>64 = 8%</td>
<td>6 = 1%</td>
<td>12 = 2%</td>
</tr>
</tbody>
</table>

Source: Baqui et al. (19).

* PTF: primary treatment failure.
Table 1 summarizes key parameters needed for the calculations presented in this report, including values specified by the trial authors (equivalence margin, power, type I error rate, sample size and the anticipated treatment failure rate used for the authors’ sample size calculations) and some important additional assumptions (prevalence of bacterial infection among neonates eligible for recruitment, sensitivity and specificity of the clinical signs used as recruitment criteria and the rate at which neonates without bacterial infection recover during the observation period).

Justifications for the assumptions in Table 1 are provided in detail below. The most important quantities that must be taken into account when computing the changes in trial design parameters are: the prevalence of bacterial infection in the population of infants eligible for trial inclusion; and the specificity and sensitivity of the recruitment criteria for bacterial infection. Together, these parameters determine the expected proportion of young infants included in the trial who have a bacterial infection, also known as the positive predictive value (PPV). The recruitment criteria are therefore discussed first, along with case-finding methods that can also impact on prevalence, sensitivity and specificity, as outlined below.

Table 1. Trial parameters and additional assumptions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specified by trial authors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equivalence margin</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Nominal power</td>
<td>90%</td>
<td>90%</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Nominal type I error rate</td>
<td>2.5%</td>
<td>2.5%</td>
<td>2.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Number randomized per trial arm&lt;sup&gt;a&lt;/sup&gt;</td>
<td>753</td>
<td>862</td>
<td>1135</td>
<td>795</td>
</tr>
<tr>
<td>Anticipated treatment failure rate among bacterial infections (used for sample size calculation)</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Additional assumptions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plausible range for prevalence of bacterial infections in neonates eligible for recruitment</td>
<td>1.5–7.5%</td>
<td>1.5–7.5%</td>
<td>1.5–7.5%</td>
<td>1.5–7.5%</td>
</tr>
<tr>
<td>Approximate sensitivity of recruitment criteria</td>
<td>70%</td>
<td>70%</td>
<td>70%</td>
<td>70%</td>
</tr>
<tr>
<td>Plausible range for specificity of recruitment criteria</td>
<td>60–80%</td>
<td>60–80%</td>
<td>60–80%</td>
<td>60–80%</td>
</tr>
<tr>
<td>Positive predictive value (% of children in trial with bacterial infection)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3–22%</td>
<td>3–22%</td>
<td>3–22%</td>
<td>3–22%</td>
</tr>
<tr>
<td>Apparent treatment failure rate among infants without bacterial infections</td>
<td>12%</td>
<td>7%</td>
<td>21%</td>
<td>9%</td>
</tr>
</tbody>
</table>

<sup>a</sup>Where sample sizes differed slightly across trial arms, the largest arm is shown.

<sup>b</sup>Calculated from the prevalence, sensitivity and specificity in the above rows.
Inclusion/exclusion criteria and case-finding

All four trials used a combination of clinical signs and symptoms to identify neonates and young infants with possible serious bacterial infection who were therefore eligible for recruitment. These signs and symptoms were based on those established by the IMCI studies (3–5). Table 2 describes trial inclusion and exclusion criteria in more detail.

All trials used a combination of active and passive case-finding. Active case-finding involved community health workers or nurses systematically searching for cases of infection. The nurses typically visited children at home looking for clinical signs, and usually visited the same child on multiple occasions. The protocols also allowed passive case-finding: for example, children whose parents presented them to a clinic with concerns about severe infection were eligible for inclusion.

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mir et al. (9)</td>
<td>- age: 0–59 days</td>
<td>- if the family agreed to admission</td>
</tr>
<tr>
<td></td>
<td>- living in the catchment area</td>
<td>- weight at presentation was less than 1 500 g</td>
</tr>
<tr>
<td></td>
<td>- refusal by family to be admitted to hospital</td>
<td>- major congenital malformations or suspected chromosomal abnormalities were present</td>
</tr>
<tr>
<td></td>
<td>- one or more signs of clinical severe infection</td>
<td>- surgical conditions that needed hospital referral</td>
</tr>
<tr>
<td></td>
<td>- had one or more signs of critical illness:</td>
<td>- they had been admitted for illness in the past two weeks</td>
</tr>
<tr>
<td></td>
<td>• movement only when stimulated</td>
<td>- they had been included previously in the study</td>
</tr>
<tr>
<td></td>
<td>• not feeding well on observation</td>
<td>- they had one or more signs of critical illness:</td>
</tr>
<tr>
<td></td>
<td>• temperature ≥ 38 °C or &lt; 35.5 °C</td>
<td>• unconsciousness</td>
</tr>
<tr>
<td></td>
<td>• severe chest indrawing (inward chest movement with every breath in one minute)</td>
<td>• convulsions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• inability to feed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• apnoea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• inability to cry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• cyanosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• bulging fontanelle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• major congenital malformations inhibiting oral antibiotic intake</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• active bleeding needing transfusion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• persistent vomiting, defined as vomiting after three attempts to feed the infant within 30 minutes, with the infant vomiting after each attempt</td>
</tr>
</tbody>
</table>
### Study 2
Tshefu et al. (11)

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>- age 0–59 days, any sign of clinical severe infection (stopped feeding well [defined as poor feeding on observation], movement only when stimulated, severe chest indrawing, and axillary temperature ≥ 38.0 °C or &lt; 35.5 °C)</td>
<td></td>
</tr>
<tr>
<td>- parents did not accept or could not access referral-level care</td>
<td></td>
</tr>
<tr>
<td>- parents gave consent to participate in the study</td>
<td></td>
</tr>
<tr>
<td>- critically ill, characterized by the presence of any of the following signs:</td>
<td></td>
</tr>
<tr>
<td>• unconsciousness</td>
<td></td>
</tr>
<tr>
<td>• convulsions</td>
<td></td>
</tr>
<tr>
<td>• unable to feed at all</td>
<td></td>
</tr>
<tr>
<td>• apnoea</td>
<td></td>
</tr>
<tr>
<td>• unable to cry</td>
<td></td>
</tr>
<tr>
<td>• cyanosis</td>
<td></td>
</tr>
<tr>
<td>• dehydration</td>
<td></td>
</tr>
<tr>
<td>• bulging fontanelle</td>
<td></td>
</tr>
<tr>
<td>• major congenital malformations inhibiting oral antibiotic intake</td>
<td></td>
</tr>
<tr>
<td>• active bleeding requiring transfusion</td>
<td></td>
</tr>
<tr>
<td>• surgical conditions needing hospital referral</td>
<td></td>
</tr>
<tr>
<td>• persistent vomiting, defined as vomiting after three attempts to feed the baby within 30 minutes</td>
<td></td>
</tr>
<tr>
<td>- very low weight (&lt; 1 500 g at the time of presentation)</td>
<td></td>
</tr>
<tr>
<td>- hospital admission for illness in the past two weeks</td>
<td></td>
</tr>
<tr>
<td>- previous enrolment in the study</td>
<td></td>
</tr>
</tbody>
</table>

### Study 3
Tshefu et al. (10)

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>- age 0–59 days</td>
<td></td>
</tr>
<tr>
<td>- fast breathing (defined as respiratory rate of ≥ 60 breaths per minute)</td>
<td></td>
</tr>
<tr>
<td>- parents did not accept or could not access referral-level care and</td>
<td></td>
</tr>
<tr>
<td>- parents gave consent to participate in the study</td>
<td></td>
</tr>
<tr>
<td>- signs of clinical severe infection defined as:</td>
<td></td>
</tr>
<tr>
<td>• poor feeding on observation</td>
<td></td>
</tr>
<tr>
<td>• movement only when stimulated</td>
<td></td>
</tr>
<tr>
<td>• severe chest indrawing</td>
<td></td>
</tr>
<tr>
<td>• axillary temperature ≥ 38.0 °C or &lt; 35.5 °C</td>
<td></td>
</tr>
<tr>
<td>- critical illness characterized by</td>
<td></td>
</tr>
<tr>
<td>• presence of unconsciousness</td>
<td></td>
</tr>
<tr>
<td>• convulsions</td>
<td></td>
</tr>
<tr>
<td>• inability to feed at all</td>
<td></td>
</tr>
<tr>
<td>• apnoea</td>
<td></td>
</tr>
<tr>
<td>• inability to cry</td>
<td></td>
</tr>
<tr>
<td>• cyanosis</td>
<td></td>
</tr>
<tr>
<td>• dehydration</td>
<td></td>
</tr>
<tr>
<td>• bulging fontanelle</td>
<td></td>
</tr>
<tr>
<td>• major congenital malformations that inhibited oral antibiotic intake</td>
<td></td>
</tr>
<tr>
<td>• active bleeding that necessitated transfusion</td>
<td></td>
</tr>
<tr>
<td>• surgical conditions needing hospital referral</td>
<td></td>
</tr>
<tr>
<td>• persistent vomiting, defined as vomiting after three attempts to feed the baby within 30 minutes</td>
<td></td>
</tr>
<tr>
<td>• very low weight (&lt; 1 500 g at the time of presentation)</td>
<td></td>
</tr>
<tr>
<td>• hospital admission for illness in the past two weeks</td>
<td></td>
</tr>
<tr>
<td>• previous enrolment in the study</td>
<td></td>
</tr>
</tbody>
</table>
### Inclusion criteria

- age 0–59 days
- residence within a predefined geographical area based on feasibility of follow-up visits
- the presence of at least one of five clinical signs of severe infection:
  - severe lower chest wall indrawing
  - axillary temperature 38.0 °C or more (≥ 100.4 °F) confirmed by a second reading
  - axillary temperature of 35.5 °C or less (≤ 95.9 °F) confirmed by a second reading
  - lethargy, defined as movement only on stimulation by the examining physician
  - history of feeding problems, confirmed by poor suck on examination

### Exclusion criteria

- Infants with fast breathing alone (respiratory rate ≥ 60 breaths per minute) were excluded
- infants with signs of critical illnesses were excluded
- presence of any of the following signs:
  - unconsciousness
  - history or presence of convulsions at assessment
  - inability to feed
  - apnoea
  - inability to cry
  - cyanosis
  - bulging fontanelle
  - major congenital malformations
  - major bleeding
  - surgical conditions needing hospital referral
  - persistent vomiting after three attempts to feed the baby within 30 minutes
  - physician’s suspicion of meningitis
- weight less than 1 500 g
- had been admitted to hospital for illness in the past two weeks
- had previously been included in the study

### Prevalence of bacterial infection in infants eligible for recruitment

The prevalence of bacterial infection in the population of young infants eligible for recruitment into each trial has to be taken into account when analysing trial designs and computing the changes in trial design parameters. As discussed above, since the trials reviewed here allowed both active and passive case-finding, this prevalence will be impacted both by true prevalence in the community and prevalence among ill infants presenting for care. Data that allow direct estimation of this prevalence are not available. Instead, a plausible range of values will be developed based on published data on the incidence of severe bacterial infection in neonates in the trial countries. This incidence has not previously been systematically investigated.

A literature review aimed to enable estimation of the incidence of serious bacterial infections through two approaches: identifying blood culture-positive bacterial infections in south Asia and in individual countries (Bangladesh, India, Kenya, Nepal and Sri Lanka (Table 3)); and reviewing studies that calculated estimates on neonatal bacterial infection.
Table 3. Incidence of culture-positive neonatal sepsis based on literature review

<table>
<thead>
<tr>
<th>Country</th>
<th>Incidence of culture-positive sepsis per 1 000 live births</th>
<th>Case fatality rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Asia</td>
<td>1.6 (Saha et al. [20])</td>
<td>NA</td>
</tr>
<tr>
<td>South Asia</td>
<td>15.7 (pooled incidence of 15 reports)</td>
<td>34.4 (Chaurasia et al. [21])</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>3 (Darmstadt et al. [22])</td>
<td>NA</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>NA</td>
<td>19.1 (Chaurasia et al. [21])</td>
</tr>
<tr>
<td>India</td>
<td>6.7 (Panigrahi et al. [23])</td>
<td>NA</td>
</tr>
<tr>
<td>India</td>
<td>16 (n = 14 studies) (Chaurasia et al. [21])</td>
<td>34.4 (Panigrahi et al. [23])</td>
</tr>
<tr>
<td>Kenya</td>
<td>5.5 (Chaurasia et al. [21])</td>
<td>NA</td>
</tr>
<tr>
<td>Nepal</td>
<td>11.6 (Chaurasia et al. [21])</td>
<td>64.7 (Darmstadt et al. [22])</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>13.6 (Chaurasia et al. [21])</td>
<td>NA</td>
</tr>
<tr>
<td>Pakistan</td>
<td>NA</td>
<td>30.9 (Panigrahi et al. [23])</td>
</tr>
</tbody>
</table>

NA: not available.

The reported incidence of culture-positive bacterial infection varied widely between countries, but also between studies, ranging from 3–16 per 1000 live births. Two publications provided estimates of identified bacterial infections. In the study of Seale et al., a systematic literature review was undertaken to obtain estimates of the incidence of possible serious bacterial infection (pSBI). They estimated a pooled pSBI incidence of 7.6 per 100 live births (24). As this is not an incidence of bacterial infection, but only that of children with clinical signs, and the signs have limited specificity, this estimate does not reflect a bacteriological correlate and is therefore quite inflated. It therefore was not included in the calculations below.

Saha et al. estimated the cause-specific incidence per 1000 live births based on results of a partially latent class model. They calculated the mean incidence of bacterial infection to be 13.2 per 1000 live births in the overall population (20). As this is a population-based incidence, it is probably higher than the point prevalence in children presenting with illness to a health facility and certainly higher than in those at any given time during a home visit. Consequently, based on the reviews of culture-positive bacterial infections and estimates provided by Seale et al. and Saha et al., a rather wide and high plausible range for the point prevalence of bacterial infections of children eligible for recruitment in the trials of 1.5–7.5%, was chosen.
Sensitivity and specificity of recruitment criteria

A perfect diagnostic test would have 100% sensitivity and 100% specificity, but this is never possible in practice. Trade-offs between sensitivity and specificity must be made: algorithms can be adjusted to be more sensitive but less specific, or more specific but less sensitive. For clinical practice, not missing a case is considered more important, so the balance is often on the side of sensitivity. Clinical signs that are too unspecific, however, lead to overtreatment and over-referral, which presents a burden to patients and health systems. This was observed, as outlined in the introduction, with the first young-infants algorithm published in IMCI in 1998 (6).

The situation with treatment trials is radically different. Specificity is of the utmost importance to demonstrate a difference between treatments. This was discussed in the design of vaccine trials, where having a specific outcome definition was considered important (25). In contrast to the two clinical IMCI studies referred to above, from which estimates of sensitivity and specificity of clinical signs were derived (sensitivity of 85% and specificity of 75%), all four trials reviewed here recruited neonates and young infants using a combination of active and passive case-finding. The active case-finding often involved repeat home visits to the same children and is therefore likely to have led to further reduced specificity compared to that estimated by the IMCI studies.

To the authors’ knowledge, there are no published studies reporting the sensitivity and specificity of clinical signs with a combination of active and passive case-finding. As a result, a wide range of possible values for the specificity of trial recruitment criteria, between 60% and 80%, was considered. For the sensitivity, which has less of an impact on power to detect a difference between treatments, a moderate value of 70% was chosen.

Apparent treatment-failure rate among infants without bacterial infections

When infants without bacterial infections are present in a trial, the rate at which their illness does not resolve during the observation period will affect the power of the trial. These infants will be recorded as treatment failures in each trial; this rate will therefore be referred to as the “apparent treatment-failure rate” among infants without bacterial infections. This value was assumed to be approximately equal to the overall treatment-failure rate, averaged across all arms. This is a reasonable approximation, as the vast majority of infants recruited will not have bacterial infections as a result of the low underlying prevalence of bacterial infection and poor specificity of recruitment criteria.

Statistical methods

For each of the four trials described above, five statistical quantities corrected for the presence of neonates and young infants without bacterial infections in the study samples were computed. These calculations are based on the following principle.

Suppose $t$ is the true difference in the proportion of treatment failures for neonates and young infants receiving the standard antibiotic regimen compared to those receiving an experimental regimen. Also suppose that $p$ is the probability that an infant recruited to the trial truly has a
bacterial infection. Assume that infants without bacterial infections will recover at the same rate when receiving either treatment. Then the difference in failure rates among all neonates and young infants included in the trial will be smaller than the true treatment difference among neonates and young infants with bacterial infections, making the treatment regimens appear closer to equivalent (13,15). Specifically, the expected treatment difference in the trial will be $p^*t$, which is less than $t$ since $p < 1$.

The following quantities were computed: detailed derivations for all calculations are provided in Annex 1. Here, these quantities and their interpretation will be described briefly. Readers interested in the statistical details should refer to Annex 1.

First, this report aims to demonstrate the impact of the presence of neonates and young infants without bacterial infections on key measures of statistical performance. For each trial, three corrected quantities were computed: type I error rate; power; and an estimate of the treatment difference, derived via maximum likelihood, and a 95% confidence interval (CI).

Secondly, changes in trial design parameters needed to demonstrate equivalence were computed after correcting for the presence of infants without bacterial infections. Specifically, for each trial the following quantities were computed: the sample size needed to achieve the study’s nominal power (typically 90%) with current recruitment criteria; and the specificity needed to achieve the nominal power at the current sample size and with a sensitivity of 70%. Specificity was emphasized over sensitivity because the specificity dominates the proportion of so-called false positives recruited (infants without bacterial infections). Power therefore is much more responsive to specificity than sensitivity.

The above calculations were based on assumptions about the prevalence of bacterial infections in the study population, the sensitivity and specificity of recruitment criteria, and the spontaneous recovery rate for infants without bacterial infections (see Annex 1 for details). The results are sensitive to these assumptions. These assumptions were therefore varied within plausible ranges (see Table 1) and the corresponding range of results was computed. A web application was also developed to allow readers to repeat all calculations under different assumptions.¹

All analyses were performed in R Version 3.6.1 (The R Foundation, 2009).²

¹ This can be accessed at: https://emgthomas.shinyapps.io/table3_calcs/
² Code to reproduce all analyses is available at: https://github.com/emgthomas/statistical_aspects_equivalence_trials
Results

Corrected type I error rate, equivalence margin, and prevalence estimates

After correcting for the presence of neonates and young infants without bacterial infections, the type I error rate for all four trials considered was around 62% at minimum (Table 4, row 1). Put another way, even when the true treatment difference among neonates and young infants with bacterial infections was clinically meaningful – at least as large as the stated equivalence margin – the risk of erroneously concluding equivalence was estimated to be 62% in the best-case scenario, and could be as high as 98%. These type I error rates are unacceptable by any standard: the nominal type I error rate for all trials was only 2.5%, and 5% is the norm (that is, a p-value of 0.05).

A simple approach to account for the presence of neonates and young infants without bacterial infections in the statistical analysis of these trials is to adjust the equivalence margin accordingly (see Annex 1). However, all four trials were drastically underpowered to detect equivalence within the adjusted margin, with power in the range of 3–14% (Table 4, row 2).

An alternative approach is to correct the estimated treatment difference and its CI so that they represent estimates of the treatment difference in young infants with bacterial infections only. This results in very wide CIs that include values well outside the trials’ stated equivalence margins (Table 4, row 3). For example, for the SATT trial (9), a conservative corrected 95% CI for the treatment difference is −21% to 5%, which includes the equivalence margin of 5%. Under different assumptions about the likely fraction of infants in the trial with true bacterial infection, however, the CI for this trial could be as large as −94% to 75%. With a CI this wide, the conclusion of equivalence is untenable. Results for the other trials were similar.

Sample size and recruitment criteria accuracy needed to show equivalence in neonates and young infants with bacterial infections

When using current recruitment criteria, enormous sample sizes of at least 12 000 per trial arm, and possibly more than 1 million, would be needed to demonstrate treatment equivalence in neonates and young infants with bacterial infection (Table 4, row 4). These sample sizes are not feasible. An alternative would be to use recruitment criteria with higher accuracy to detect bacterial infection while maintaining a more reasonable sample size. Table 4 (row 5) shows that in a trial with 5000 infants per arm, recruitment criteria with a specificity of at least 96% for bacterial infection would be needed (assuming the sensitivity was 70%; note that lower sensitivity would necessitate even higher specificity).
Table 4. Trial features and results corrected for the presence of subjects with non-bacterial infections

<table>
<thead>
<tr>
<th>Corrected parameter</th>
<th>Study 1 Mir et al. (9)</th>
<th>Study 2 Tshefu et al. (11)</th>
<th>Study 3 Tshefu et al. (10)</th>
<th>Study 4 Baqui et al. (19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Type I error (%)</td>
<td>64–83</td>
<td>86–98</td>
<td>62–82</td>
<td>74–92</td>
</tr>
<tr>
<td>3. Estimate and 95% CI for treatment difference</td>
<td>$-0.09$(\text{(-0.21, 0.05)})</td>
<td>$-0.09$(\text{(-0.18, 0.02)})</td>
<td>$-0.12$(\text{(-0.27, 0.04)})</td>
<td>$-0.07$(\text{(-0.19, 0.06)})</td>
</tr>
<tr>
<td>4. Required sample size to achieve nominal power and type I error rate (to the nearest 1 000)</td>
<td>17 600–1 312 000</td>
<td>12 000–820 000</td>
<td>21 000–545 000</td>
<td>14 000–1 024 000</td>
</tr>
<tr>
<td>5. Specificity required to achieve nominal power and type I error rate for sample size per trial arm of 5 000 and sensitivity of 0.70</td>
<td>0.966–0.994</td>
<td>0.963–0.993</td>
<td>0.969–0.994</td>
<td>0.993–0.965</td>
</tr>
</tbody>
</table>

- Treatment difference: treatment failure rate in experimental arm minus treatment failure rate in standard arm. If the trial had more than two arms, results are shown comparing one experimental arm only to standard treatment, as specified in footnotes.
- Standard treatment: intramuscular procaine benzylpenicillin and gentamicin for seven days; experimental treatment: oral amoxicillin and intramuscular gentamicin for seven days.
- Standard treatment: intramuscular procaine benzylpenicillin and gentamicin for seven days; experimental treatment: oral amoxicillin for seven days.
- Standard treatment: intramuscular procaine benzylpenicillin and gentamicin for seven days; experimental treatment: intramuscular gentamicin and oral amoxicillin for seven days.

Visualization for power and sample size correction

Fig. 2 visualizes the impact of the percentage of neonates and young infants without bacterial infection on power and sample size required to establish treatment equivalence. It shows an example based on the Mir et al. (9) trial of sepsis in young infants in Pakistan. First, if the trial recruitment criteria had perfect specificity (100%) for bacterial infection – meaning the PPV is 1, so that all infants had culture-positive bacterial infections at recruitment – a sample size of 757 per trial arm would be required to detect treatment equivalence within a margin of 5% and with 90% power. This is close to the required sample size of 750 computed by Mir et al. (9) and is illustrated on the plot by the lowest horizontal (green) line corresponding to \(n = 757\), which meets the 90% power contour when \(PPV = 1\).

\(3\) Additional assumptions: type I error rate = 2.5%; treatment-failure rate among infants with bacterial infection = 10%; treatment-failure rate among infants without bacterial infection = 12%.
Fig. 2 also demonstrates what happens when the trial recruitment criteria have less than perfect accuracy for bacterial infection. Assume the sensitivity of these criteria is 70% and the specificity is between 60% and 80%. Based on a literature review, the prevalence of neonatal sepsis among Pakistani young infants screened for the trial could be as low as 1.5% or as high as 7.5% (Table 3). These assumptions imply a PPV of between 2.6% and 22.0%, represented on the plot by the two vertical orange lines (see Annex 1 for PPV formula). At the originally computed sample size of 757, this means that the true power of the study to detect equivalence is between 3% and 10%, drastically lower than the assumed 90%. This is illustrated by the power contours at the intersections of the horizontal green line (sample size n = 757) and the two vertical orange lines (PPV = 2.6% and PPV = 26.0%).

Finally, follow the vertical orange lines upwards until they meet the desired power contour of 90%. Moving horizontally across from here to the Y-axis shows that the sample size required to detect equivalence with 90% power is between 17 585 and 1 312 447 per trial arm, between 23 and 1700 times larger than the originally calculated n = 757.

**Fig. 2. Impact of PPV on study power and required sample size**

*Note:* the plot shows how power, shown by the contours, rapidly varies as a function of PPV (PPV: percentage of recruited infants with true bacterial infection, shown on the X-axis) and sample size per trial arm (Y-axis). Note that both axes are shown on a log scale. As an example, the assumptions for this plot mirror that of Study 1 ([9]; see Table 1): sensitivity = 70%; specificity = 60–80%; prevalence of bacterial infection = 1.5–7.5%; desired power = 90%; type I error rate = 2.5%; equivalence margin = 5%; treatment-failure rate among infants with bacterial infections = 10%; treatment-failure rate among infants without bacterial infections = 12% (Table 1). An interactive version of Fig. 2 that allows the user to vary these assumptions is available at: https://emgthomas.shinyapps.io/power_calc/.
Interactive tools for computing corrected measures

An online interface\(^4\) is available for repeating all calculations shown above. An interactive version\(^5\) of Fig. 2 is also available. These tools are intended to allow the reader to vary the assumptions shown in Table 1 to observe the impact on the results and to obtain results for other equivalence trials.

\(^4\) This can be accessed at: https://emgthomas.shinyapps.io/table3_calcs/
\(^5\) Available at: https://emgthomas.shinyapps.io/power_calc/
Discussion

This report shows that the conclusions of treatment equivalence made by four recently published trials of sepsis in neonates and young infants are not tenable (9–11, 19). It became evident that these studies were critically underpowered to establish equivalence of antibiotic treatments among neonates and young infants with bacterial infection, the only infants for whom antibiotic treatments can be effective (Table 3). All four trials had an extremely high probability – 60% at minimum – of concluding treatment equivalence even if the experimental regimen was in reality worse than the standard of care by a clinically important margin of 5%. Indeed, after accounting for the likely fraction of infants in the trials who did not have bacterial infections, the CIs for the treatment difference among neonates and young infants with bacterial infection may extend well beyond the 5% equivalence margin for all trials. These corrected CIs may even include values that represent an extremely harmful effect of the experimental treatment relative to standard treatment, with treatment differences possibly higher than 20%. As such, the efficacy of the experimental antibiotic regimens studied requires urgent re-examination.

With current methods for trial recruitment, huge sample sizes of 12 000 per trial arm or more would be required to detect treatment equivalence for infants with bacterial sepsis (Table 4). Such sample sizes probably are not feasible. The alternative for conducting an equivalence trial is to recruit infants using screening tools with a higher specificity for bacterial infection. A way forward for this could be to include only neonates and young infants with positive blood culture in the study analysis. As the trials attempt to reflect a real-life situation, and results would only be available later, this would need to be done as a post hoc analysis of culture-positive cases.

Many neonates and young infants with a bacterial infection nevertheless will still have negative blood cultures, so demanding this might be too stringent. An alternative for conducting a trial would be using a combination of tests such as white cell count, differential white cell count, procalcitonin and c-reactive protein, or serological markers of bacterial infection that would, independently or in combination, increase the likelihood that test-positive infants have a bacterial infection. With a sample size of 5000 per trial arm and sensitivity of 70%, a specificity of at least 96% is needed to detect equivalence with 90% power. Although more specific screening tools and larger sample sizes will increase trial costs, the potential cost in terms of human lives put at risk by continuing to base sepsis treatment guidelines for neonates and young infants on questionable evidence is certainly much greater.

Another option may be to include a placebo arm in these trials. If the proportion of infants with bacterial infections in the trials is very low, as the analysis above suggests, then the apparent failure rate for young infants treated with antibiotics would likely appear to be very similar to placebo for all treatment arms. One recent trial in Pakistan included a placebo arm in a study of less severely ill young infants with fast breathing only, who were randomized to oral amoxicillin or placebo (26). The trial was stopped because there was a significant difference, with an approximately two-fold higher treatment failure in the placebo group (2.8% versus 5.9%). This indicates there were some infants with bacterial infection in the trial who did benefit from antibiotics. However, including a placebo arm still does not provide a means to estimate the treatment difference among children with bacterial infection, which, as demonstrated in this report, was most likely much larger than the 3% difference reported by this trial. The ethical dilemma of including a placebo arm in trials concerning treatment of serious bacterial infection in infants also emerges.
It has to be emphasized that the design of equivalence trials should reflect the prevalence of the disease as well as the sensitivity and specificity of the chosen inclusion and exclusion criteria. All four trials presented here were performed with a highly preselected study group. Based on the chosen exclusion criteria, the infants who were randomized were likely to suffer from mild disease. This observation is consistent with the low mortality rates in all trials and study arms of 1–2%, which are much lower than previously reported mortality rates for neonates and young infants with severe infections ranging from 18.0–64.7% with a pooled case fatality rate of 34.4 (20,24). The preselection of such a study population, as pointed out in this report, is surprising, given that the intention of the trials was to evaluate treatment for neonates and young infants who are ill enough to require hospital admission under normal circumstances. One AFRINEST trial even excluded severely ill children as part of the study protocol.

While the authors of the studies explained the calculation of the sample size in detail, the number of participants in each trial arm was based on presumed prevalence of infection, and the power of the study and sample sizes were not revised or critically assessed once pre-trial assumptions were found to be not applicable.

This report reviewed four large trials cited in the guidelines on treatment of sepsis in young infants when referral is not possible (12). One additional study cited in the guidelines was not considered because it included only 150 infants per trial arm (27) compared to 750 or more for the four trials in this review. The smaller sample size of this study means the statistical limitations will be even greater compared to the trials reviewed. A systematic search was conducted to identify studies published after the guidelines were launched and which might add to the evidence base on the effectiveness of simplified antibiotic treatments administered in the community. This revealed two additional studies. The first compared oral amoxicillin to placebo arm and is discussed above (26). The second was a site-specific reanalysis of data from the multisite AFRINEST trial (28). While these studies were not reviewed in the same depth as in this report, they do not affect the conclusions drawn.

**Limitations**

This report made a number of assumptions about the possible range of true values for the underlying prevalence of bacterial infections, sensitivity and specificity of recruitment criteria, and apparent treatment-failure rates for infants without bacterial infection (Table 1). The results are sensitive to these assumptions. For prevalence, a conservative range was considered which included values that are likely to be much higher than the true value. This would have led to underestimates of the uncertainty in the treatment differences estimated by the trials. The SATT trial took blood cultures for 2067 (84%) infants in the trial, of which 1712 were negative, 273 were contaminated and 81 were culture-positive (9). Excluding the contaminated samples, this suggests about 4.5% of infants in the trial were culture-positive. Another trial of young infants in Pakistan found a similar result, with pathogens isolated from 11 (5.0%) of 218 blood cultures (27). These are likely to be underestimates of the fraction of children in these trials with bacterial infection, since prior studies suggest almost 75% of blood cultures from children with sepsis fail to recover an organism (29,30). These results therefore are roughly consistent with the assumption shown in Table 1 that between 3% and 22% of children recruited to the trials reviewed in this report had a bacterial infection.

To the authors’ knowledge, currently there are no published estimates of the sensitivity and specificity of the combination of active case-finding and clinical signs used for recruitment to
the trials. A range of conservative values for the specificity and a moderate value for the sensitivity was therefore used in this report. An online interactive tool\(^1\) is available to allow readers to vary all of the above assumptions.

Throughout this report, it is assumed that the apparent treatment-failure rate for neonates and young infants without bacterial infections is the same in all trial arms. This assumption may not hold if, for example, infants with viral infections recover faster when treated at home. If so, this would make the trial results appear even more favourable to the experimental oral antibiotic regimens, further increasing the risk of type I error.

Finally, in this report the maximum likelihood estimator of the treatment difference among neonates and young infants with bacterial infections accounting for the presence of young infants without bacterial infection in the trials was presented. As discussed in Annex 1, this maximum likelihood estimator and corresponding CI are likely to be unreliable when the PPV of the trial recruitment criteria is very small. This reflects the fact that the trial data contain little information about the true treatment difference when few infants in the trial have bacterial infections.

**Conclusions**

This report shows that several published treatment trials for possible serious bacterial infections were unable to demonstrate equivalence of the experimental treatment options with the standard of care. Developing protocols and policies based on these results is therefore not admissible.

Bacterial infections in neonates and young infants remain a serious issue requiring action by the global community to reduce associated mortality. Unfortunately, the methods used by the trials reviewed led to poor-quality evidence and do not progress this endeavour. Other avenues need to be explored, some of which are discussed in the previous papers identifying clinical signs\(^{(2,5,6)}\). Pretending to have a less demanding way of treating bacterial infections without knowing whether it is adequate, instead of investing in health systems, is ethically dubious. Efforts have to go in the direction of making sure newborns are born into, and live in, a protective environment that reduces their risk of infection and, when they have a possible bacterial infection, are treated with the best possible care in settings that improve their chances of survival.
References


Annex 1. Details of statistical calculations

Here, we (the authors) provide details of the calculations presented in the report. Our objective is to recompute key design parameters and results for equivalence trials of antibiotics to treat neonatal sepsis, after accounting for the expected proportion of infants recruited to the trials that do not have true bacterial infections.

Let:

- \( \alpha \) be the type I error chosen by the trial authors;
- \( 1 - \beta \) be the power chosen by the trial authors;
- \( \delta \) be the equivalence margin chosen by the trial authors;
- \( \eta \) be the stated sample size per trial arm, when all arms have the same sample size;
- \( \eta_s (\eta_E) \) be the fixed total number of subjects in the standard (experimental) arm, when the sample sizes differ between trial arms;
- \( \eta_{SF} (\eta_{EF}) \) be the observed number of subjects who experience treatment failure in the standard (experimental) arm;
- \( TFR_{BS} (TFR_{BE}) \) be the true treatment failure rate for study subjects with true bacterial infection under the standard (experimental) treatment;
- \( TFR_{NB} \) be the true “treatment-failure rate” (proportion who do not recover) for study subjects with non-bacterial infections – throughout, we assume this rate to be the same for both arms of the trial;
- \( t \) be the treatment difference: treatment-failure rate in standard arm minus treatment-failure rate in experimental arm;
- \( \hat{t} = \frac{n_{EF}}{n_E} - \frac{n_{SF}}{n_S} \) be the usual estimated treatment difference;
- \( \rho \) be the prevalence of bacterial infections in the population from which study subjects are recruited (eligible for screening);
- \( \theta \) be the sensitivity of the screening test used to select study subjects;
- \( \varphi \) be the specificity of the screening test, and;
- \( PPV \) be the positive predictive value of the screening test, that is, the proportion of recruited participants who have bacterial infection, calculated as \( PPV = \theta \rho / (\theta \rho + (1 - \varphi)(1 - \rho)) \).

Type 1 error

Here, we compute the true type I error rate: the probability of concluding equivalence of the two treatments when there is in fact a true treatment difference among neonates with true bacterial infection. To do this accurately, we need to account for the fraction of neonates in the
trial who have bacterial infections. Here, the null hypothesis is that the true treatment difference is exactly equal to the stated equivalence margin, $\delta$, among neonates with true bacterial infection. That is, $H_0$ is:

$$TFR_{BE} - TFR_{BS} = \delta.$$ 

Under $H_0$, the overall true treatment difference in the theoretical study population – that is, the pool of individuals who could be recruited to the trial, which includes neonates with and without bacterial infections – will be:

$$\delta_{corr} = TFR_E - TFR_S = \delta PPV$$

where

$$TFR_S = TFR_{BS} PPV + TFR_{NB}(1 - PPV)$$

is the overall treatment failure rate in the standard arm, and

$$TFR_E = TFR_{BE} PPV + TFR_{NB}(1 - PPV)$$

$$= (TFR_{BS} + \delta) PPV + TFR_{NB}(1 - PPV)$$

is the overall treatment failure rate in the experimental arm.

If a type I error rate of $\alpha$ is desired, the usual test statistic for non-inferiority trials is the upper limit of a $100(1 - 2\alpha)%$ confidence interval for the estimated treatment difference:

$$T = \hat{i} + \Phi^{-1}(1 - \alpha)\sigma$$

where

$$\sigma^2 = Var(\hat{i}) = \frac{TFR_E(1 - TFR_E)}{n_E} + \frac{TFR_S(1 - TFR_S)}{n_S}$$

and $\Phi$ is the cumulative distribution function of a standard normal random variable.

In practice, we replace $\alpha$ by a consistent estimator:

$$\hat{\sigma} = \sqrt{\frac{n_{EF}/n_E(1 - n_{EF}/n_E)}{n_E} + \frac{n_{SF}/n_S(1 - n_{SF}/n_S)}{n_S}}$$

This is the test statistic used by all trials discussed in the main report.

Assuming there are no study subjects with non-bacterial infections ($PPV = 1$), under the null hypothesis $H_0$, $\hat{i}$ is approximately normally distributed with mean $\delta$ and variance $\alpha^2$ (for large sample sizes). The rejection region for the test is $T < \delta$. First, we check that the type I error rate for this test is indeed $\alpha$. Assuming the null hypothesis is true and $PPV = 1$, the type I error rate, or probability of falsely rejecting the null and concluding treatment equivalence, is:

$$P(T < \delta | H_0, PPV = 1) = P \left( \frac{\hat{i} - \delta}{\sigma} < -\Phi^{-1}(1 - \alpha) \left| H_0, PPV = 1 \right. \right)$$

$$= P(Z < \Phi^{-1}(\alpha))$$

$$= \Phi(\Phi^{-1}(\alpha)) = \alpha$$

where $Z$ is a standard normal random variable.

If there are study subjects with bacterial infections ($PPV < 1$), but this is ignored and the same test statistic and rejection region are used, the distribution of $T$ changes: its mean is now $\delta_{corr}$. Its true variance $\alpha^2$ also changes relative to the $PPV = 1$ case, since $TFR_E$ and $TFR_S$ depend on
PPV, but the form of the estimator \( \hat{\sigma} \) doesn’t change and is still consistent for \( \hat{\sigma} \). We can now calculate the type I error rate exactly as before, but using this new distribution for \( T \). The probability of falsely concluding equivalence of the treatment for neonates with true bacterial infections will be:

\[
P(T < \delta | H_0, \text{PPV} < 1) = P \left( \frac{\hat{t} - \delta_{\text{corr}}}{\sigma} < \frac{\delta - \delta_{\text{corr}}}{\sigma} + \Phi^{-1}(\alpha) \bigg| H_0, \text{PPV} < 1 \right)
\]

\[
= P \left( Z < \frac{\delta(1 - \text{PPV})}{\sigma} + \Phi^{-1}(\alpha) \right)
\]

\[
= \Phi \left( \frac{\delta(1 - \text{PPV})}{\sigma} + \Phi^{-1}(\alpha) \right)
\]

\[=: \alpha_{\text{corr}}.\]

\( \alpha_{\text{corr}} \) is the corrected type I error: the probability of falsely concluding treatment equivalence under \( H_0 \) when the test statistic \( T \) is used but \( \text{PPV} < 1 \).

Note that when \( \text{PPV} = 1 \), \( \alpha_{\text{corr}} = \Phi(\Phi^{-1}(\alpha)) = \alpha \) as expected, and that \( \alpha_{\text{corr}} > \alpha \) whenever \( \text{PPV} < 1 \).

**Power**

Let \( \beta_{\text{corr}} \) be the type II error rate after correcting for the presence of non-bacterial infections. Thus, \( 1 - \beta_{\text{corr}} \) is the corrected power: To compute the power, we assume that there is no treatment difference among children with bacterial infections, that is, \( TFR_{\text{BE}} = TFR_{\text{BS}} \), note this also implies no overall treatment difference, that is, \( TFR_E = TFR_S \). In order to detect this treatment equivalence in a study sample that includes children with non-bacterial infections, we need to use the corrected equivalence margin \( \delta_{\text{corr}} = \delta_{\text{PPV}} \).

We therefore compute the power to detect equivalence under the corrected equivalence margin using the usual formula for power in equivalence trials:

\[
1 - \beta_{\text{corr}} = \Phi \left( \sqrt{\frac{n(\delta_{\text{PPV}})^2}{2TFR(1 - TFR)}} - \Phi^{-1}(1 - \alpha) \right),
\]

where \( TFR = TFR_E = TFR_S \) is the treatment-failure rate in both trial arms, and can be calculated as shown in the previous section.

**Corrected treatment difference and confidence interval**

For a trial that includes subjects with non-bacterial infections, we can estimate the treatment difference among children with bacterial infections only when \( \text{PPV} \) is fixed and known. As we will see, we do not need to know \( TFR_{\text{NB}} \) to obtain the maximum likelihood estimate (MLE) of the treatment difference. The likelihood is:

\[
\mathcal{L} = TFR_{\text{SF}}^{n_{\text{SF}}}(1 - TFR_S)^{n_{\text{SF}} - n_{\text{SF}}} TFR_{\text{EF}}^{n_{\text{EF}}}(1 - TFR_E)^{n_{\text{EF}} - n_{\text{EF}}}
\]

where \( TFR_S \) and \( TFR_E \) depend on \( TFR_{\text{BS}} \) and \( TFR_{\text{BE}} \) as shown in section 1. Taking the log of \( \mathcal{L} \) and finding the maximum in \( TFR_{\text{BS}} \) and \( TFR_{\text{BE}} \) subject to the restrictions \( 0 \leq TFR_{\text{BS}} \leq 1 \) and \( 0 \leq TFR_{\text{BE}} \leq 1 \), gives the following MLEs:
The estimated treatment difference among children with bacterial infections only is:

\[ \hat{t} = \hat{TFR}_{BE} - \hat{TFR}_{BS}. \]

Because of the discrete nature of this estimator, we obtain a \( (1 - \alpha)100\% \) confidence interval by simulation as follows. Let \( m \) be the number of simulations. For \( i = 1, \ldots, m \), do the following.

1. Simulate \( n_{FE} \) from \( \text{Binom}(n_E, \hat{TFR}_E) \) where
   \[ \hat{TFR}_E = \frac{n_{SF} - n_S(1 - PPV)TFR_{NB}}{n_S PPV} \]
   if \( \frac{n_{SF}}{n_S} < TFR_{NB}(1 - PPV) \),
   \[ \frac{n_{EF} - n_E(1 - PPV)TFR_{NB}}{n_E PPV} \]
   if \( TFR_{NB}(1 - PPV) \leq \frac{n_{SF}}{n_S} \leq PPV + TFR_{NB}(1 - PPV) \),
   \[ 1 \]
   if \( \frac{n_{SF}}{n_S} > PPV + TFR_{NB}(1 - PPV) \),

and

\[ \hat{TFR}_{BE} = 0 \]

2. Simulate \( n_{FS} \) from \( \text{Binom}(n_S, \hat{TFR}_S) \) where
   \[ \hat{TFR}_S = \frac{n_{EF} - n_E(1 - PPV)TFR_{NB}}{n_E PPV} \]
   if \( \frac{n_{EF}}{n_E} < TFR_{NB}(1 - PPV) \),
   \[ \frac{n_{EF} - n_E(1 - PPV)TFR_{NB}}{n_E PPV} \]
   if \( TFR_{NB}(1 - PPV) \leq \frac{n_{EF}}{n_E} \leq PPV + TFR_{NB}(1 - PPV) \),
   \[ 1 \]
   if \( \frac{n_{EF}}{n_E} > PPV + TFR_{NB}(1 - PPV) \),

We take our confidence interval to be the \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles of the \( \hat{t}_i \) that is, the interval will be \( (l, u) \) where

\[ l = \min_s \left( \frac{1}{m} \sum_{i=1}^{m} \{ \hat{t}_i \leq s \} \right) \geq \alpha/2 \]

and

\[ u = \min_s \left( \frac{1}{m} \sum_{i=1}^{m} \{ \hat{t}_i \leq s \} \right) \geq 1 - \alpha/2. \]

We now make some observations. First, when \( PPV = 1 \), the above reduces to the usual estimator for a difference in proportions.

Second, when the following condition holds

\[ TFR_{NB}(1 - PPV) \leq \frac{n_{SF}}{n_S}, \frac{n_{EF}}{n_E} \leq PPV + TFR_{NB}(1 - PPV), \]

we have

\[ \hat{t} = \hat{TFR}_{BE} - \hat{TFR}_{BS} = \left( \frac{1}{PPV} \right) \left( \frac{n_{EF}}{n_E} - \frac{n_{SF}}{n_S} \right). \]
The standard error of this estimator is
\[
s.e.(\hat{t}) = \frac{1}{\text{PPV}} \left( \frac{TFR_S (1 - TFR_S)}{n_S} + \frac{TFR_E (1 - TFR_E)}{n_E} \right)^{1/2}
\]
which can be estimated as
\[
\hat{s.e.}(\hat{t}) = \frac{1}{\text{PPV}} \left( \frac{TFR_S (1 - TFR_S)}{n_S} + \frac{TFR_E (1 - TFR_E)}{n_E} \right)^{1/2}
\]
where
\[
\hat{TFR}_S = \hat{TFR}_{BS}\text{PPV} + TFR_{NB}(1 - \text{PPV}) = \frac{n_{SF}}{n_S} \quad \text{and}
\]
\[
\hat{TFR}_E = \hat{TFR}_{BE}\text{PPV} + TFR_{NB}(1 - \text{PPV}) = \frac{n_{EF}}{n_E}.
\]
In this case, when \(\text{PPV} = 1\), both \(\hat{t}\) and \(\hat{s.e.}(\hat{t})\) reduce to the usual estimates for a difference in proportions for independent groups and its standard error, as expected. Further, \(\hat{t}\) does not depend on \(TFR_{NB}\), which makes intuitive sense as any difference between the treatment arms must be due to differences among subjects with bacterial infections only. However, \(\hat{s.e.}(\hat{t})\) depends on both \(\text{PPV}\) and \(TFR_{NB}\) because \(TFR_{NB}\) contributes to the variance in \(n_{SF}\) and \(n_{EF}\).

On the other hand, \(\hat{s.e.}(\hat{t})\) does not depend on \(TFR_{NB}\) because the variance in \(n_{SF}\) and \(n_{EF}\) can be captured entirely through the observed data.

When the sample size is large enough, a 95% confidence interval for \(\hat{t}\) relying on the usual asymptotic normal approximation is \(\hat{t} \pm 1.96\hat{s.e.}(\hat{t})\). The estimator and confidence interval are equal to the usual estimator for a difference in proportions and its confidence interval (ignoring subjects with non-bacterial infections) multiplied by \(1/\text{PPV}\). The usual test statistic in an equivalence trial is the upper limit of the confidence interval \(\hat{t} \pm 1.96\hat{s.e.}(\hat{t})\), and the rejection region is \([-1, \delta]\). Thus we note that for large enough sample sizes, this approach is equivalent to computing the treatment difference and confidence interval as usual (ignoring subjects with non-bacterial infections), but using the corrected equivalence margin \(\delta_{corr} = \delta_{PPV}\). The latter is the approach we took when computing power, type I error, and sample size.

Relying on the normal approximation may result in confidence intervals that include values outside the possible range \([-1,1]\); this occurs partly because the normal approximation is only valid for large enough sample sizes. It is for this reason that we compute confidence intervals by simulation.

Finally, the MLE and simulated confidence interval have limitations. The MLE \(\hat{t}\) is consistent for the true treatment difference \(t\), but may have poor performance in small samples. In particular, the definition of a “small sample” depends on the value of \(\text{PPV}\) as well as \(n_S\) and \(n_E\): if \(\text{PPV}\) is very small, the effective sample size will be small in that the sample will contain little information about \(t\). Similarly, for small \(\text{PPV}\), the variance of \(\hat{t}\) may be large. In addition, when \(\text{PPV}\) is small, the discrete nature of the simulated confidence intervals may lead to poor coverage.
Required sample size

Let \( n_{\text{corr}} \) be the sample size required to detect equivalence when \( t = 0 \) after correcting for the presence of non-bacterial infections. As for the power and type I error calculations above, this requires us to correct the equivalence margin to \( \delta_{\text{corr}} = PPV \delta \). The required sample size can then be computed using the standard formula for equivalence trials:

\[
n_{\text{corr}} = \frac{2TFR(1 - TFR)(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2}{(PPV \delta)^2}
\]

where

\[
TFR = TFR_B PPV + TFR_{NB}(1 - PPV) \text{ and } TFR_B = TFR_{BS} = TFR_{BE}.
\]

Required sensitivity and specificity

We aim to find values of the sensitivity \( \theta \) and specificity \( \phi \), if any exist, that would allow us to detect equivalence at a given sample size \( n \) with the nominal power \( 1 - \beta \) and type I error rate \( \alpha \). The first step is to solve Equation 1 in the previous section for \( PPV \). We then solve \( PPV \) for \( \theta \) and \( \phi \).

Equation 1 is quadratic in \( PPV \). Solving for \( PPV \) gives

\[
PPV = \frac{-b \pm \sqrt{\Delta}}{2a}
\]

where

\[
a = -(A + (TFR_B - TFR_{NB}))^2,
b = (1 - 2TFR_{NB})(TFR_B - TFR_{NB}), \quad \text{and}
c = TFR_{NB}(1 - TFR_{NB}),
\]

\[
\Delta = b^2 - 4ac, \quad \text{and}
A = \frac{n\delta^2}{2(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2}.
\]

Note that since, for all values of the parameters, \( \alpha < 0 \) and \( c \geq 0 \), we will always have \( \delta > 0 \). Thus this quadratic equation always has at least one solution. Typically, we will have only one solution in the range \([0, 1]\).

We then solve the equation

\[
PPV = \theta p / (\theta p + (1 - \phi)(1 - p))
\]

for \( \theta \) and \( \phi \). This gives the linear relationship

\[
\phi = 1 - \frac{p(1 - PPV)}{1 - p} \theta
\]

and provides an infinite set of solutions \((\phi, \theta)\). As expected, to achieve the required \( PPV \), there is a trade-off between sensitivity and specificity: a higher sensitivity means a lower specificity is required, and vice versa.
The WHO Regional Office for Europe

The World Health Organization (WHO) is a specialized agency of the United Nations created in 1948 with the primary responsibility for international health matters and public health. The WHO Regional Office for Europe is one of six regional offices throughout the world, each with its own programme geared to the particular health conditions of the countries it serves.

Member States

Albania
Andorra
Armenia
Austria
Azerbaijan
Belarus
Belgium
Bosnia and Herzegovina
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Georgia
Germany
Greece
Hungary
Iceland
Ireland
Israel
Italy
Kazakhstan
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Malta
Monaco
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Republic of Moldova
Romania
Russian Federation
San Marino
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Tajikistan
Turkey
Turkmenistan
Ukraine
United Kingdom
Uzbekistan