Recommendations to FDA for Interim Endpoints for Clinical Trials in Community-acquired Bacterial Pneumonia

Foundation for the National Institutes of Health Biomarkers Consortium Project Team

CABP Docket ID: FDA-2009-D-0136

Contents

0 Executive Summary

During recent decades, the efficacy endpoints for Community-Acquired Bacterial Pneumonia (CABP) registrational studies relied on a clinical assessment of cure requiring "complete resolution of signs and symptoms" based on a combination of non-standardized physician-based observations and comments collected from the patient by the physician as well as on the investigator's assessment of the need for alternative antibiotic therapy. As non-inferiority clinical trial design advanced during the late 20th and early 21st century, it became apparent to the FDA and others that the development of more readily quantifiable, reproducible, and externally verifiable endpoints would improve the design of present-day non-inferiority clinical trials for CABP.

In developing updated approaches to endpoints, it was also recognized that outcome measures used for studies that support drug registration for CABP must be relevant for clinical practice. Although the level of detail and accuracy in measurement needed in the setting of clinical trials may differ from that needed in clinical practice, a description of the pivotal (Phase 3, or registrational) clinical trials as conducted is an integral part of the prescribing information and must be based directly on the trial data as collected and analyzed. The choice of primary endpoint for a trial may thus need to balance a variety of competing demands.

In parallel discussions of the design of studies for skin infections, the idea arose that standardized assessments of patient response in CABP in the first few days of therapy might provide key insights into both drug effect and options for trial design (Food and Drug Administration 2010). Consequently, and at the request of FDA, in early May, 2010, the Foundation for the National Institutes of Health (FNIH) convened a Project Team for a Biomarkers Consortium Project entitled "Developing Endpoints for Clinical Trials of Drugs for Treatment of Acute Bacterial Skin and Skin Structure Infections and Community-Acquired Bacterial Pneumonia (Phases 1 and 2). The Project Team membership included broad participation from NIH, FDA, the academic research community (including members of the Infectious Diseases Society of America (IDSA)), and interested biopharmaceutical companies.

This document summarizes the work of the Biomarkers Consortium Project Team. Over a series of meetings the group reviewed the available historical and modern data and found that progressive improvement in four symptoms (cough, dyspnea, chest pain, and sputum production) during the first 4 days of therapy was sufficiently well documented that an early response endpoint measure could be proposed. To assess durability of response and other late events, supportive information should be obtained by assessing outcomes at a fixed timepoint after therapy has been completed. Such information could include a late response endpoint similar to the traditional test-of-cure endpoint. Although based on limited data and requiring further research, an early response endpoint can be used to anchor a non-inferiority trial for this indication. The early response endpoint is thus suggested for possible use by FDA in review of registrational trials and approval of applications in CABP while further research into this area is conducted.

1 Introduction/ Background

1.1 Background

Long known as the "Captain of the Men of Death," Community-Acquired Bacterial Pneumonia (CABP) is a well-recognized and frequent syndrome (Spellberg, Talbot et al. 2008). Pneumonia remains the sixth leading cause of death in the United States and the number one cause of infectious disease-related death. Mortality rates in the pre-antibiotic era were often substantial (e.g., rates $\geq$ 60% were reported in patients $\geq$ 60 years of age) (Spellberg, Talbot et al. 2008), and even higher rates were reported for subsets such as patients with bacteremia (Fleming and Powers 2008). With the availability of effective antibiotics and advances in supportive care, mortality rates in the antibiotic era are reduced but still substantial at 10–20% (Ochoa-Gondar, Vila-Corcoles et al. 2008).

Over the past decades, CABP has often been a key element of the initial registration indication(s) for new agents. Based on early observations that fever (core body temperature elevated above the normal range) in particular tended to resolve in just a few days with adequate therapy (Petersdorf, Cluff et al. 1957; el Moussaoui, Opmeer et al. 2006) vs. an average of 8-10 days (Osler 1910; Bullowa 1937) in the pre-antibiotic era, resolution of fever (elevated core body temperature) as well as the more gradual resolution of pulmonary symptoms was used in many early reports as the basis for judging adequate efficacy. As subsequent antibiotics were introduced, trials relied on a clinical assessment of cure that required "complete resolution of signs and symptoms" based on a combination of non-standardized physician-based observations and comments collected from the patient by the physician as well as on the investigator's assessment of the need for alternative antibiotic therapy.

Approaches to endpoints in CABP that reduce dependence on physician-based observations or patient-based reporting have been considered but have to date been frustrated by practical issues. Mortality could be used as an endpoint in trials of CABP (Fleming and Powers 2008) (Spellberg, Fleming et al. 2008) and overall population mortality (10-20%) is theoretically high enough to support this approach (Ochoa-Gondar, Vila-Corcoles et al. 2008). However, the observed overall mortality rate includes patients who cannot be enrolled (e.g., those who died on or before hospital admission). As a result, the mortality rate of the enrolled patient population in recent trials has been $\leq$ 5%, a figure that is too low to make this endpoint practical (Pertel, Bernardo et al. 2008; Tanaseanu, Bergallo et al. 2008; Tanaseanu, Milutinovic et al. 2009). Placebo-controlled superiority-based designs are also not possible in the study of CABP because of the dramatic mortality and morbidity benefit of antibiotic treatment (Spellberg, Fleming et al. 2008).

Thus, development of new agents for this indication has always relied on active-controlled non-inferiority studies using a clinical assessment of cure. As trials based on this approach have detected inferior agents (Pertel, Bernardo et al. 2008)) and as future trials will of necessity rely on comparative agents approved using this approach, a draft FDA Guidance for non-inferiority studies of CABP in which continued use of this approach was proposed in 2009 (Food and Drug Administration 2009, March).

Recent discussions regarding non-inferiority study design have, however, recognized the importance of improving the design of non-inferiority clinical trials for this indication. First, the "clinical response" endpoints used in prior CABP trials have depended upon a physician-based assessment and may also have included biomarkers that are not on the causal pathway of the disease. The composite of these various measures was left to clinician discretion. The concern is that this approach does not meet the regulatory criterion that endpoints must be "well-defined and reliable." Endpoints must be either direct measures of how a patient functions, feels or survives or properly validated replacement endpoints for such measures in the appropriate context of use. In an effort to improve the strength of evidence when efficacy is evaluated in non-inferiority trial designs, work was thus undertaken to assess the clinical relevance of various endpoints, to better define those endpoints, and as well as to evaluate the optimal timing for the assessment of efficacy in patients with CABP.

An additional particular focus for review was to provide strong estimates of treatment effect size relative to placebo therapy based on well-defined and reliable measures derived as closely as possible from patient-based information and taken at specific points in time. Having reliable estimates of treatment effect size is essential for a non-inferiority trial design. Although the historical evidence outlined above is consistent with a large effect, the available data are limited in that:

1. The endpoints used in the historical trials do not specifically define the variables measured and the reliability of how they are measured, two fundamental components of endpoints for pivotal trials.
2. The data are incomplete and cannot be audited.
3. The data are taken from studies conducted many years ago, so their relevance to the modern clinical setting could be questioned. Since the time of these reports, there have been many changes in medical therapy such as improvements in supportive care and ready availability of antipyretics or anti-inflammatory agents which may alter treatment effects on biomarkers such as body temperature.
4. The data are not well controlled for severity of illness (or its potential to become severe) or baseline predictors of outcomes.
5. Development of biomarkers for use in chronic infections (Micheel, Ball et al. 2010) has led to the recognition that the biomarkers commonly used in acute infection should be evaluated carefully to ensure good linkage to underlying syndrome and evaluation and qualification of their use when used as outcome measures in clinical trials. Although both general biomarkers (core body temperature, heart rate) and disease-specific biomarkers (respiratory rate in pneumonia, erythema in skin infections) demonstrate supportive temporality and consistency, they are consequences of the infection rather than causes of the infection.
6. The data do not provide direct access to patient-based outcomes similar to those used in patient-reported outcome (PRO) tools.

Physicians and patients have a natural interest in the overall outcome at the end of therapy and thereafter — the goal is resolution of the infection, no relapse, no late sequelae, and no significant adverse effects of the therapy itself. The traditional clinical trial Test-of-Cure (TOC) endpoint taken at a time after therapy has completed has had the goal of capturing all of these

elements, but in so doing it has incorporated a subjective decision-making element that makes it ill-defined from a regulatory perspective. Recognition of this potential ambiguity is useful for understanding the role that a novel regulatory endpoint might play in resolution of this problem.

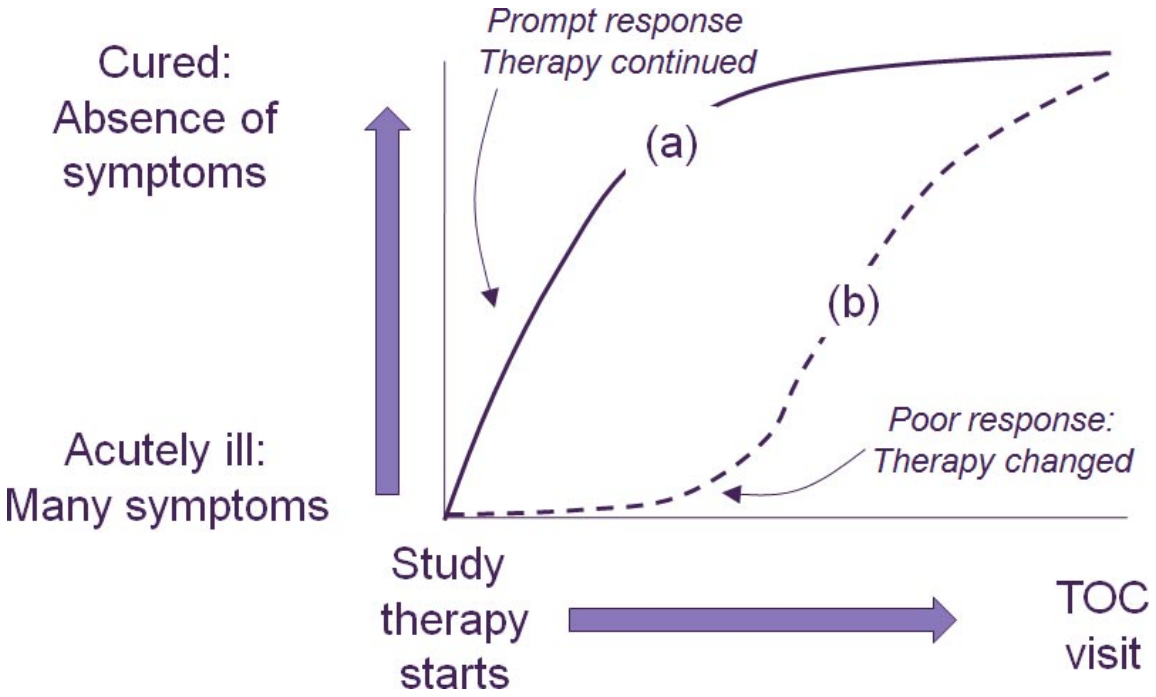
Specifically,

1) The final patient state associated with the *traditional* TOC endpoint of "Cured" was characterized by the complete or near complete absence of symptoms associated with the infection and the return of relevant physiological parameters to normal (or premorbid status). Acceptance of a "near complete" absence of symptoms was justified in part by prior studies that showed that complete return to previous baseline status in CABP may take months, which is longer than the time point at which TOC measures have been obtained (Metlay, Atlas et al. 1998). While clinicians often express confidence in their ability to reliably define and measure near complete absence of symptoms in the setting of clinical practice and thus often consider such an endpoint to be well-defined when taken at a sufficiently late point in time, measures of improvement need to be clearly defined and quantified in the setting of clinical trials.

2) Contributing reasons to the traditional TOC endpoint becoming ill-defined are the incorporation of components that either are not well-defined and reliable or are biomarkers where effects on these measures have not been shown to reliably predict effects on direct measures of how a patient feels, functions, or survives, and the inclusion of events that occur before the TOC endpoint. Although viewed as relevant by patients and physicians, such earlier events contain some subjective decision-making components.
    a) The decision to continue or discontinue study drug therapy, especially during the first few days of therapy.
    b) The decision to utilize salvage therapy.
    c) The observation (or not) of therapy-limiting adverse events.

3) Thus, the patient's state <u>alone</u> at the late time point of the traditional TOC endpoint may not be sensitive to study drug effects. As illustrated in Figure 1, both patients (a) and (b) could be judged as Cured at the TOC visit, but they reach this state in different ways:

Figure 1. Similar outcomes at a traditional late TOC visit, but different courses



New insights around the events that occur early in therapy and the possibility of a new early endpoint can contribute to addressing these problems. The work described in this document provides the basis for a consistent and objective description and documentation of the key early decision-making steps, thereby creating a well-defined approach to endpoints that capture and describe the overall effectiveness of study drug therapy (initial efficacy, sustained efficacy, and tolerability).

1.2 Approach Taken by the Project Team

At the request of FDA, in early May 2010 FNIH convened a Project Team with broad participation from NIH, FDA, the academic research community (including members of the Infectious Diseases Society of America (IDSA)), and interested biopharmaceutical companies to address the issues described above. The group has worked to develop a consensus on alternative primary and secondary endpoints that might improve the quality of future clinical trials of CABP.

In developing updated approaches to endpoints, it was also recognized that outcome measures used for studies that support drug registration for CABP must be relevant for clinical practice. Although the level of detail and accuracy in measurement needed in the setting of clinical trials may differ from that needed in clinical practice, a description of the registrational (Phase 3) clinical trials as conducted is an integral part of the prescribing information and must be based directly on the trial data as collected and analyzed. The choice of primary endpoint for a trial may thus need to balance a variety of competing demands.

It was thus agreed that the approach to developing such endpoints would involve two steps.

First, available data would be used to develop a set of interim recommendations that would permit sponsors to continue development of drugs for this indication (and, consequently, that FDA would consider data based on these recommendations as pivotal data for review of applications for marketing authorization) (Phase 1). Second, a series of investigations would be undertaken into further possible improvements of endpoint measures and/or development of new measures (Phase 2). Such work might well benefit from incorporation of clinical trial data obtained using the interim recommendation endpoints as a starting point. Thus, improvements of endpoint measures would become available in the near future. The recommendations presented here are interim; they are based on currently available evidence, but there is an urgent need for further research to address the gaps in research elucidated during the Project Team's review.

This initiative is particularly important at a time when the incidence of treatment-resistant pathogens is increasing (Boucher, Talbot et al. 2009). The recent slowdown in antimicrobial drug development and lack of clarity regarding regulatory requirements for registration of these important drugs adds further urgency to this undertaking.

2 Summary of Project Team Process

The members of the Project Team convened for a series of meetings during 2010 and 2011. Over the course of these meetings, the group discussed the historical literature, recent publications, and data from several available modern clinical studies. The group developed a consensus on a two-phase process to identify primary and secondary endpoints for ABSSSI (discussed separately) and CABP (this document).

2.1 Phase 1: Retrospective Data Analyses

The goal of this phase was to perform retrospective analyses of datasets from existing clinical studies to a) refine/confirm currently proposed outcome measures by determining how they performed in a modern clinical trial setting; b) help identify additional endpoints or biomarkers that might be relevant. The Project Team has identified several sources of data from existing modern industry clinical trials that have been used as an in-kind contribution to the project.

These analyses, which have also been contributed in-kind to the project, have been based in each case on a statistical analysis plan (SAP) drafted by qualified biostatisticians who are part of the Project Team; each SAP was shared with the entire Project Team for comment and approval prior to initiating the analyses.

2.1.1 Summary of Existing Datasets

1. Historical data
    a. Bullowa, J. G. M. (1937). Chapter II. The course, symptoms and physical findings. The management of pneumonias. New York, NY, Oxford University Press: 36-76.
    b. Finland, M., W. C. Spring, et al. (1940). Immunological Studies on Patients with Pneumococcic Pneumonia Treated with Sulfapyridine. J Clin Invest 19(1): 179-99.

c. Flippin, H. F., J. S. Lockwood, et al. (1939). The treatment of pneumococcic pneumonia with sulfapyridine. JAMA-J Am Med Assn 112: 529-534.
d. Meakins, J. C. and F. R. Hanson (1939). The treatment of pneumococcic pneumonia with sulfapyridine. Can Med Assoc J 40: 333–6.
e. Osler, W. (1910). Specific infectious diseases: Lobar pneumonia. The Principles and Practice of Medicine. New York, D. Appleton and Company: 164-192.
f. Wilson, A. T., H. A. Spreen, et al. (1939). Sulfapyridine in the Treatment of Pneumonia in Infancy and Childhood. JAMA 112: 1435-1439.
g. Summary analyses of early antibiotic era data (Presentation by Mary Singer, 8 Dec 2009 FDA AIDAC, available online at www.fda.gov).

2. Pfizer Pharmaceuticals generously provided the primary data tables from the clinical trials which the two comparative studies of tigecycline vs. levofloxacin which underpinned tigecycline's approval for CABP:
    a. Tanaseanu, C., C. Bergallo, et al. (2008). Integrated results of 2 phase 3 studies comparing tigecycline and levofloxacin in community-acquired pneumonia. Diagn Microbiol Infect Dis 61(3): 329-38.
    b. Tanaseanu, C., S. Milutinovic, et al. (2009). Efficacy and safety of tigecycline versus levofloxacin for community-acquired pneumonia. BMC Pulm Med 9: 44.

3. Cubist Pharmaceuticals generously provided analyses of responses over time in the ceftriaxone arm from a study of daptomycin vs. ceftriaxone for CABP:
    a. Pertel, P. E., P. Bernardo, et al. (2008). Effects of prior effective therapy on the efficacy of daptomycin and ceftriaxone for the treatment of community-acquired pneumonia. Clin Infect Dis 46(8): 1142-1151.

4. Both FDA and Cerexa, Inc. generously provided analyses from the two studies of ceftaroline vs. ceftriaxone which underpinned ceftaroline's approval for CABP:
    a. FDA Briefing document for 7 Sep 2010 AIDAC: Ceftaroline Fosamil for the Treatment of Community-acquired Bacterial Pneumonia and Complicated Skin and Skin Structure Infections. Available online at www.fda.gov
    b. Cerexa Briefing document for 7 Sep 2010 AIDAC: Ceftaroline Fosamil for the Treatment of Community-acquired Bacterial Pneumonia and Complicated Skin and Skin Structure Infections. Available online at www.fda.gov

2.1.2 Review of Historical Data

A review of the course of untreated pneumonia provided a useful baseline against which to judge the clinical course of the disease in the modern era and also from which to draw insights into possible endpoints (see material summarized in Section 5.1). Reviews of work by Osler (Osler 1910) and Bullowa (Bullowa 1937) provided illustrations of the typical course of symptoms associated with the syndrome of acute bacterial pneumonia, including cough, dyspnea, chest pain especially worsened with coughing, and expectoration of sputum. The patient would experience a steady deterioration during the early course of disease with progressive respiratory symptoms and change in mental status. If the patient survived, the initial sign of resolution would be a

drenching sweat after the eighth or ninth day (the "crisis"). Initial resolution was followed by onset of suppurative complications in some patients.

Based on these data, a critical analysis of the course of illness in the untreated patient can be generated. Early in the course of illness, the untreated patient has fever (elevated core body temperature) and multiple respiratory symptoms. Prior-generation physicians wrote more about elevated body temperature because it was so obvious and because the day of the "crisis" was such an important clinical event. But, it is also clear that respiratory symptoms were prominent and progressive and that they also began to improve once the fever began to resolve. Osler describes this transition well when he writes, "After persisting for seven to ten days, the crisis occurs, and with a fall in the temperature the patient passes from the condition of extreme distress and anxiety to one of comparative comfort." It is thus well documented that in the untreated patient, respiratory symptoms were not improved by day 3-4 but rather that steady deterioration could occur during this period.

These results were contrasted with the experience in the early antibiotic era.[1] Based on data from the early antibiotic era (Flippin, Lockwood et al. 1939; Meakins and Hanson 1939; Wilson, Spreen et al. 1939; Finland, Spring et al. 1940), an antibacterial treatment effect using clinical recovery as an endpoint can be described. As described by early investigators in qualitative terms, the effect was rapid and striking (Flippin, Lockwood et al. 1939): "From the very beginning of this study, we have been impressed, as were Evans and Gaisford (1939), by the striking frequency with which the initiation of drug treatment was followed within 24 hours or less by a critical drop in the patient's temperature. This temperature drop was not immediately accompanied by any significant changes in lung signs but always reflected a marked improvement in the toxemia and the general well being of the patient. Resolution of the pneumonia then followed within a variable period of days".

Using an endpoint characterized by a general improvement in the patient's clinical condition as observed and recorded by the physician, substantial treatment effects can be estimated from these data:

- A quantitative estimate of treatment effect for symptom resolution at 48 to 72 hours is 29% (95% confidence interval = 21-37%) (Finland, Spring et al. 1940).
- A quantitative estimate of treatment effect for clinical recovery at day 3 is 72% to 77% (Bullowa 1937; Flippin, Lockwood et al. 1939; Meakins and Hanson 1939).
- Quantitative estimates of treatment effect for mean days to clinical improvement, fall in temperature, and clinical recovery were 2.5, 3.4 and 4.2 days, respectively (Wilson, Spreen et al. 1939).

Although these data suggest a significant effect of antibacterial agents, the data also have a number of limitations:

- The data are mostly observational or from small studies.
- Cross-study comparisons were used to determine treatment effect.
- The endpoints not clearly defined, but were clinically reasonable.

[1] Data adapted from a presentation by Mary Singer, 8 Dec 2009 FDA Anti-Infective Drugs Advisory Committee.

But, the studies also have counterbalancing strengths (they were contemporaneous; except for Finland's 1940 study (Finland, Spring et al. 1940), mortality rates ranged from 3-7% in treated patients and were thus similar to mortality rates reported in contemporaneous controlled studies; the data were primarily from cases of pneumococcal disease). Taken together, this collection of pre-antibiotic and early antibiotic era data suggested a significant treatment effect at approximately day 3–4 after initiation of therapy. On this basis, an exploratory, hypothesis-generating analysis was undertaken of the tigecycline-levofloxacin CABP dataset in an effort to better define the variables measured in the "clinical response" endpoint.

2.1.3 Tigecycline vs. Levofloxacin - Hypothesis Generation

In this phase of the work, data from the two pivotal trials underpinning the registration of tigecycline for CABP were analyzed (Tanaseanu, Bergallo et al. 2008; Tanaseanu, Milutinovic et al. 2009). These studies enrolled patients with an average age of 51 years with a distribution of PORT scores (I-V, microbiologic modified ITT [intention-to-treat] population) of 22%, 31%, 27%, 19%, and 1%. In both studies, tigecycline was compared with levofloxacin as monotherapy for CABP. Patient-level data on the time course of four symptoms were available for analysis. Specifically, scores of absent, mild, moderate, or severe had been recorded for each of these four symptoms:

a. Cough
b. Pleuritic chest pain
c. Dyspnea
d. Sputum production

Based on the idea that rapid symptom improvement might be expected early in the course of therapy (but not necessarily complete resolution over such a short period of time), a series of exploratory initial analyses focused on three possible definitions of response. The first two definitions sought to define the concept of "some symptom better with no other worse," whereas the third definition measures disappearance of all symptoms:

a. The first day when some baseline symptom was better, with none of the other symptoms having become any worse.
b. The first day when some baseline symptom was now absent, with none of other symptoms having become any worse.
c. The first day when all symptoms were reported to be absent.

Further analyses explored two types of "temporary responders", that is, patients with initial response who did not maintain that response. Such patients were defined as either:

a. Patients with a response at Study Day[2] 3, 4, or 5, but with failure to maintain that response at all later times.

Or

[2] Throughout this document, Study Day 1 corresponds to the day of initiation of study therapy. An observation on Study Day 2 (usually the next calendar day) would be taken approximately 24h after initiation of therapy, an observation on Study Day 3 would be taken approximately 48h after therapy initiation, Study Day 4 would be approximately 72h after therapy initiation, and Study Day 5 would be approximately 120h after therapy initiation. The datasets discussed in this paper did not rigidly define specific time windows but rather appear to have followed a largely calendar-day based convention.

b. Patients with a response at Day 3, 4, or 5, but with a failure to maintain the response at the TOC visit.

Finally, the Project Team considered the possibility that endpoints with a stricter response definition might either reduce the problem of "temporary response" or offer a usefully different pattern of response over time. Thus, variant definitions of response along the scale of absent, mild, moderate or severe were also considered:

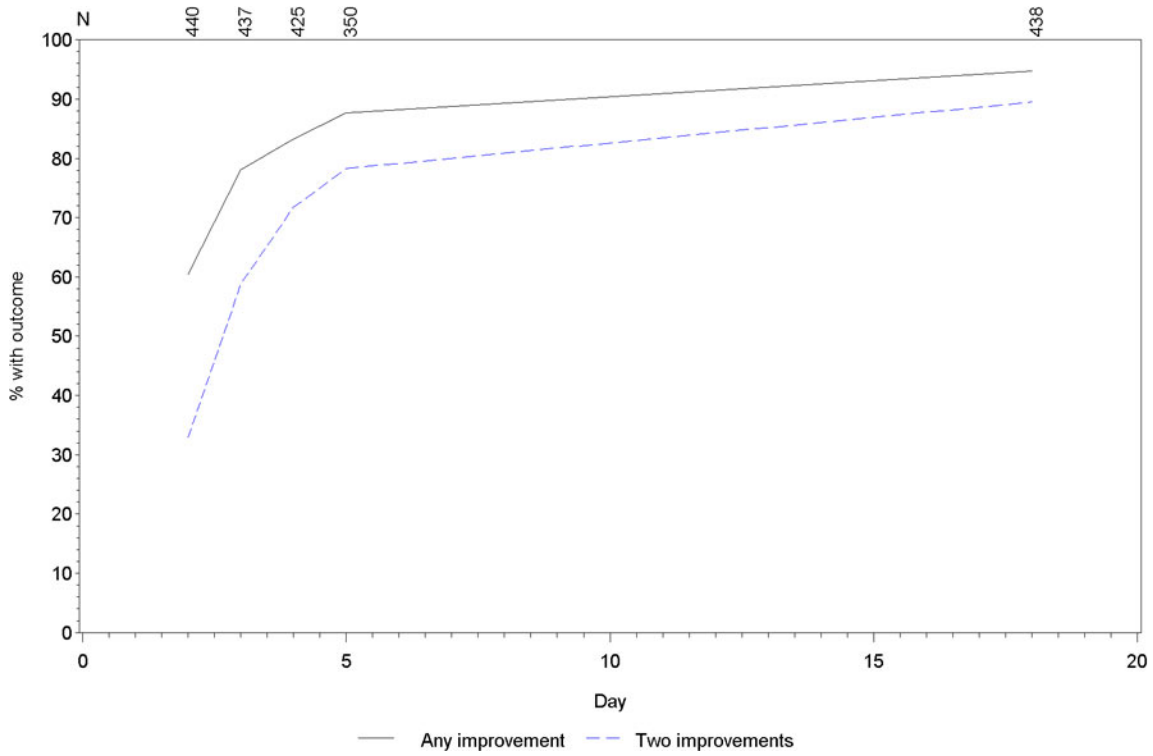
a. Any improvement from baseline in 2 of 4 symptoms, with none of other symptoms having become any worse.
b. A 2-point improvement (e.g. from severe to mild or moderate to absent) in one symptom, with none of other symptoms having become any worse.
c. A 2-point improvement in one symptom, a 1-point improvement (e.g. from severe to moderate or mild to absent) in another symptom, with none of other symptoms having become any worse.

These additional observations were relevant to understanding the available data:

a. Most patients had daily observations and measurements during the first 4 Days of therapy.
b. Subsequently, significant time gaps would span observations and measurements. As the exact Day of a change could not be estimated, missing observations were not replaced by last observation carried forward.
c. Most patients have a TOC and Late Follow-up (FU) data point, but these observations did not occur on the same Day for all patients. Thus, the number of observations on specific Days after about Day 5 becomes quite variable.
d. The highest number of observations was on Days 1-4, at the TOC visit, and at the FU visit.
e. Baseline findings for at least two symptoms (that is, a score of Mild, Moderate, or Severe rather than a score of Absent) were present in 96% of patients (See supplemental data in Section 5.2, Table 6) and thus most patients could be judged to improve based on a two-symptom rule. In addition, 93% of patients had a sufficient number of symptoms to meet a response rule requiring a 2-point change in at least one symptom and 91% had sufficient symptoms to meet a response rule requiring a 2-point change in one symptom accompanied by a 1-point change in another symptom.
f. Although the strength of symptom scores of Absent, Mild, Moderate, and Severe is limited by the lack of well-validated definitions, the Project Team believes that the perception that drives a change in category for an individual patient is likely to reflect a meaningful change in patient status. Further, the short duration of illness is likely to permit reasonable recall.

The core results are shown in Figure 2. In this graph, the y-axis shows the percentage of subjects meeting rules in which response meant improvement in one symptom by one point (solid line) or in two symptoms by one point (dashed line) with no worsening of any other symptom. As can be seen, rapid improvement can be documented during the first five Study Days based on analyses of these symptoms. This result appears similar to the qualitative descriptions of clinical response in the early antibiotic era literature.

Figure 2: Improvement in CAP symptoms over Days 1-5, all patients



To aid with understanding how the new definitions performed in the context of the traditional TOC endpoint, supplemental analyses were performed for the subsets judged Cured vs. Failed at the traditional TOC endpoint (Section 5.2, Figure 4 and Figure 5). These analyses have a number of limitations (principally, they rely on the traditional TOC and its subjective elements which the Project Team seeks to avoid), but they proved useful during consideration (see below) of the choice of rule and time point that produced the least number of both temporary responses and responses that were discordant with the traditional endpoint.

The definition that was determined to offer the greatest merit was the one which required improvement in at least two symptoms, each by at least one point (that is, an improvement by one category such as from Moderate to Mild). As noted above, 96% of patients had sufficient baseline symptoms to permit them to meet the response criterion and the Project Team decided that improvement-in-two-symptom-categories supported a larger treatment effect that would correspond to clinically meaningful effects. Conclusions based on this definition were:

a) The requirement for improvement by at least one point in two symptoms yields improvement rates of 59, 72, and 78% on Days 3, 4, and 5 for all patients combined (Figure 2).

b) Agreement between the response at Days 3–5 with clinical cure/failure as judged at the traditional TOC visit was assessed as a guide to maximizing sensitivity to early treatment effect while also limiting the number of temporary responders (those showing an initial response meeting the rule but with subsequent worsening of symptoms):

i) Broadly, earlier times (Day 3) were better for limiting the number of patients who are an early improver, but are ultimately classified to be a clinical failure at the TOC visit. This conclusion appears biologically plausible but the data on this point are limited by varying numbers of observations on each Day in this small dataset.
ii) Similarly and with the same limitations, later times (Day 5) were better for limiting the number of patients who are an early non-improver, but who were ultimately judged as a clinical cure at the TOC visit.
iii) Reasoning that failing to predict ultimate successful outcome is a lesser error than incorrectly predicting success, earlier times (Days 3-4) overall seemed to offer the best balance.

c) In summary, minimizing the number of patients who improve early and then are not improved later was determined as most appropriate and is facilitated in this dataset by an early evaluation at Days 3 and 4:
i) Improvement rates for ultimate clinical failures were minimized on these Days.
ii) Evaluation on these days minimized the number of patients showing improvement at this time, but not subsequently classed as an improvement.

Two alternatives to the definition requiring a one-point improvement in at least two symptoms were analyzed in parallel, with all three rules shown in Table 1. The Project Team also analyzed the frequency with which an initial response was not sustained as judged by failure to meet the same rule at the TOC visit (Table 2).

Table 1: Response rates for three possible response definitions

Day	One-point improvement in two symptoms	Two-point improvement in one symptom	Two-point improvement in symptom, one-point improvement in another symptom
Day 3	257/437 (59%)	155/437 (35%)	138/437 (32%)
Day 4	305/425 (72%)	209/425 (49%)	193/425 (45%)
Day 5	274/350 (78%)	203/350 (58%)	192/350 (55%)

Table 2: Rates of temporary improvement - Met the response rule at an early time point but not at TOC

Day	One-point improvement in two symptoms	Two-point improvement in one symptom	Two-point improvement in symptom, one-point improvement in another symptom
Day 3	10/251 (4%)	7/149 (5%)	7/134 (5%)
Day 4	8/300 (3%)	8/204 (4%)	7/190 (4%)
Day 5	10/271 (4%)	4/199 (2%)	6/190 (3%)

Overall, the definition of early response requiring a one-point improvement in at least two symptoms overall appeared most consistent with both TOC data and the prior descriptions of antibiotic response. Despite the above-discussed limitations of the TOC endpoint, the Project Team felt that given these evaluations looked specifically at symptom improvement, considering

the correlation of early and late response was an appropriate way to use all of the available information to calibrate the proposed early endpoint rules.

The Project Team discussed at length the merits of the alternative rules. Although the greater stringency of the two-point improvement endpoints might offer greater sensitivity to treatment effects, some members of the Project Team thought these were more difficult to interpret. First, they were concerned that it is not clear what a two-point change means. Second, response rates based on a two-point improvement were lower than seemed clinically plausible. Finally, the one-point improvement concept is similar to the idea of "Any improvement" and might be considered a simpler definition to understand and use.

Finally, all three endpoints had relatively low (<5%) and similar rates of temporary improvement. Discordance rates for only those patients successful at TOC were similar to those above for all patients combined. Discordance rates in those unsuccessful at TOC were difficult to interpret due to low numbers — the absolute number of discordant patients in this group was low.

Based on these discussions, the Project Team concluded that the one-point-improvement-in-two-symptoms rule was a reasonable approach but that alternative rules could be (re)considered and developed as additional data become available.

Conclusions From This Analysis

1) This analysis of the three definitions and the reliability of the early response measure suggest that a one-point improvement in two symptoms at Day 3, 4, or 5 should be the focus of further analysis when symptoms are classified on a four-point scale consisting of absent, mild, moderate, severe.
2) Alternative approaches were possible but were considered to present obstacles that were greater than those posed by the consensus endpoint definition.
3) The two-point improvement definitions included the issue that defining a two-point change is a more challenging hurdle to meet and all two-point improvements may not be equal. Although all one-point improvements may as well not be equal, a one-point improvement could be taken as a meaningful step from the patient's perspective and a pair of such improvements for two different symptoms was likely a strong finding. Future research is needed to better define responder criteria.
4) In addition, response rates based on a two-point improvement were in a range (ca. 50%) that would require substantially larger clinical trial sample sizes than those required for response rates closer to 70–80%. The Project Team discussed the possibility of selecting the rule based on its impact on sample size but could only conclude that further research was required on that point.
5) "Any improvement" could be considered a simpler definition to understand and use.
6) All three definitions of the endpoints had relatively low and similar rates of temporary improvement, so the choice of definition was not a factor in discordance.
7) In terms of the timing of the outcome assessment, Days 4 and 5 had higher response rates with similar levels of discordant responses.
8) No data are available on the content validity or reliability of the scale used in this analysis; however, the analysis of the presented data in this form was used to understand the disease

pattern. Future research would help to evaluate the content validity, understandability to patients and reliability of scales.

a) Although scale reliability had not been validated, it was noted that the short duration of the illness would facilitate accurate day-to-day comparison by the patient and that change in rating level was likely to reflect the course of the illness. The validity of this assumption should be studied in future research.
b) Some members of the Project Team noted that although the use of this scale was appropriate, the terminology for Mild, Moderate, Severe, and Absent needed better and more precise definition.

9) The available data indicated that most patients with CABP receiving effective therapy demonstrate a two-symptom improvement, each by at least one point.

2.1.4 Analysis of Ceftriaxone Treatment Data - Limited Hypothesis Testing

Using the ideas developed from investigation of the tigecycline-levofloxacin analysis, an analysis plan was developed for the ceftriaxone data from the ceftriaxone-daptomycin CABP trial (Pertel, Bernardo et al. 2008):

In brief, two CABP studies were conducted with daptomycin (Cubicin, formerly Cidecin) vs. ceftriaxone. Of these, the first was completed in 2000 and the data from the ceftriaxone arm were generously made available for this analysis. The second study was stopped when the results of the first study's results became available.

In this study, the mean age of the enrolled patients in the ITT ceftriaxone group was 56 years with a PORT Risk Class distribution (I-V) of 0%, 44%, 30%, 27%, and 0% (Pertel, Bernardo et al. 2008). The same four symptoms as previously analyzed (cough, chest pain, dyspnea, and sputum production) were serially recorded for each patient. A weakness of this dataset is that symptoms are only recorded as present or absent. A further weakness is the small number of failures in the ceftriaxone arm. Thus, the exploratory analysis provides only limited hypothesis testing. Although the definition used in the study protocol and presented to investigators was to evaluate "improvement" in symptoms, the case report forms did not conform with this definition since investigators were only offered the choices of "present" or "absent" for each symptom.

Of the evaluable population of 286 patients, 97.6% had two or more symptoms at baseline. Overall, 81.1% of subjects had at least one symptom resolve by Day 4 and 58.1% had at least two symptoms resolve by Day 5 (Table 3). Similar to prior observations (Metlay, Fine et al. 1997), cough took longer to resolve than other symptoms. For example, in the subset of patients with at least one symptom eradicated, only 30% had cough resolved by Day 5, whereas 60, 52, and 66% of patients have resolution of dyspnea, chest pain, and sputum production, respectively (Table 4). A similar pattern was observed in the subset with recorded resolution of at least two symptoms (Table 5).

Table 3. Number of symptoms resolved by Study Day

Study Day	Resolution of at least one symptom	Resolution of at least two symptoms
3	193/286 (67.5%)	67/279 (24.0%)
4	232/286 (81.1%)	127/279 (45.5%)
5	243/286 (85.0%)	162/279 (58.1%)

Table 4. Timing of resolution of at least one symptom

Study Day	N with at least ONE symptom eradicated	Cough eradicated (%)[a]	Dyspnea eradicated (%)[a]	Chest pain eradicated (%)[a]	Sputum production eradicated (%)[a]
3	193/286 (67.5%)	18 (9)	91 (47)	75 (39)	103 (53)
4	232/286 (81.1%)	44 (19)	126 (54)	111 (48)	143 (62)
5	243/286 (85.0%)	73 (30)	145 (60)	126 (52)	161 (66)

[a]Data in these columns show n eradicating the given symptom / N eradicating at least one symptom (%)

Table 5. Timing of resolution of at least two symptoms

Study Day	N with at least TWO symptoms eradicated	Cough eradicated (%)[a]	Dyspnea eradicated (%)[a]	Chest pain eradicated (%)[a]	Sputum production eradicated (%)[a]
3	67/279 (24.0%)	17 (25)	50 (75)	47 (70)	47 (70)
4	127/279 (45.5%)	42 (33)	91 (72)	85 (67)	97 (77)
5	162/279 (58.1%)	71 (44)	114 (70)	106 (65)	128 (79)

[a]Data in these columns show n eradicating the given symptom / N eradicating at least one symptom (%).

Relationship between Symptom Resolution and Clinical Outcome. The sensitivity and the specificity of at least one symptom vs. two symptoms resolved were assessed. To evaluate sensitivity, the cure rates and the percentage of patients who had at least two symptoms resolved were of interest. Of those classified as a cure at the TOC visit, 82% had at least one symptom resolved by Day 4. However, 82% of those classified as a failure at TOC likewise had at least one symptom resolve. On the other hand, 62% of the patients judged to be a cure at TOC had at least two symptoms resolved on study Day 5 vs. only 18% of subjects ultimately judged to be a failure. Once again, such analyses must be interpreted with caution since "cure at TOC" is used as the "gold standard" in such comparisons, even though it has not been established to be a validated surrogate endpoint for long-term resolution of symptoms.

Characteristics of Four Failed Patients with at Least Two Symptoms Resolved. The patients who were classified as a failure but had at least two symptoms resolved were evaluated more closely. Three of the failed patients each had a persistence or progression of radiographic abnormalities (a pre-specified "failure" definition) at TOC. Patient 1 improved over time and symptoms resolved from Days 3–5, but the patient had persistence or progression of radiographic abnormalities at the TOC visit. Patient 2 had sporadic improvement and a normal chest radiograph at TOC. Patient 3 had chest pain and cough that were resolved at Day 5 but came

back at the TOC; this patient also had persistence or progression of radiographic abnormalities. These discrepancies should not be over-interpreted and could have been due to worsening of a baseline symptom or the presence of symptoms outside those recorded. The final patient showed resolution according both to the study definition on study Days 3–5 and based on the symptom data at TOC follow-up; there was no clear reason why this person failed in the disposition data set. However, this patient had a medley of other problems and was taking several concomitant medications including some potentially effective antibiotics. As for the first three patients, other symptoms could have worsened or been present at baseline and resulted in the failure classification.

Conclusions From Review of the Ceftriaxone Dataset

- The data are limited by recording of only present/absent for each symptom and do not correspond to the study protocol's definitions for improvement.
- The number of patients with symptoms of interest at baseline is similar to what was observed previously: 98% of patients had two or more symptoms at baseline.
- Symptoms at baseline were similar for patients who were classified as cure or failure.
- With the caveats noted above regarding the meaning of the TOC assessment, one-symptom resolution did not correlate well with an assessment at the TOC visit. On the other hand, two-symptom resolution had a broad, general agreement with the TOC assessment and with the analysis of the levofloxacin-tigecycline dataset.
- Three of four failures (75%) who did show resolution of two or more symptoms on Day 5 had persistence or progression of radiographic abnormalities.
- Forty percent of patients who were an investigator-determined cure at TOC did not have two or more symptoms resolved from baseline by Study Day 5. In particular, cough was noted to be a persistent symptom that did not resolve completely with antibiotic therapy during the usual observation period.
- Overall, these findings are consistent with the observation from the tigecycline-levofloxacin data set that improvement of two or more symptoms on approximately Day 4 of therapy (approximately 72h into the course of therapy) is indicative of response to therapy.

2.1.5 Analyses Undertaken During Review of the Ceftaroline Phase 3 CABP Studies

The FDA has recently reviewed two phase 3 non-inferiority trials compared ceftaroline with ceftriaxone in the treatment of adults with CABP and on the basis approved ceftaroline for this indication. Enrolled subjects had mean age of 61 years with 62% of the subjects in PORT category III and 38% in PORT category IV.

As noted in the FDA-approved prescribing information, "To evaluate the treatment effect of ceftaroline, an analysis was conducted in CABP patients for whom the treatment effect of antibacterial agents may be supported by historical evidence. The analysis endpoint required subjects to meet signs and symptoms criteria at Day 4 of therapy: A responder had to both (a) be in stable condition according to consensus treatment guidelines of the Infectious Diseases Society of America and American Thoracic Society, based on temperature, heart rate, respiratory rate, blood pressure, oxygen saturation, and mental status (Mandell, Wunderink et al. 2007); (b)

show improvement from baseline on at least one symptom of cough, dyspnea, pleuritic chest pain, or sputum production, while not worsening on any of these four symptoms."

The response rates at study Day 4 for microbiologically evaluable patients were 69.6% and 69.0% for ceftaroline and 58.3 and 61.4% for ceftriaxone for trials 1 and 2, respectively.

The FDA reviewers also suggested that knowing whether the clinician was assessing stability based on the above-noted definition at that earlier time point could help in the evaluation of efficacy as a measure in addition to symptoms and evaluated separately (not as a composite outcome measure). FDA evaluated the literature that IDSA/ATS has published on the criteria for establishing stability. These objective criteria for stability (body temperature ≤ 37.8 °C, pulse $\leq$ 100 beats per minute, respiratory rate ≤ 24 breaths per minute, stable blood pressure ≥ 90 mm Hg, oxygen saturation $\geq 90\%$, and normal mental status) have been suggested as a means to help clinicians understand when it is appropriate to discharge a patient from the hospital. Although only one element of this definition is directly tied to how a patient feels or functions (normal mental status), the FDA view parallels the practical clinical sense that these measurements are directly tied to the historical data on response and can serve to support a non-inferiority margin. The quantitative relationship between biomarkers and symptoms is an area that needs further research, as correlations between biomarkers and outcomes of how patients feel, function and survive may represent a useful starting point but are insufficient to evaluate and qualify biomarkers as outcome measures.

As also stated in the FDA-approved prescribing information, FDA concluded that the historical data available at the time of this drug's review were insufficient to establish the magnitude of the drug effect for antibacterial drugs using clinical response at the TOC time point. However, the FDA review team determined that the product label should provide a full description of the entire course of treatment for CABP. The protocol-specified analyses in the CABP trials included the clinical cure rate at the test of cure (TOC) visit (8–15 days after treatment ended).

Conclusions From Review of the Ceftaroline US FDA CABP Registration Dataset

- A recent drug registration has been based on a response definition based on (a) achieving clinical stability based on temperature, heart rate, respiratory rate, blood pressure, oxygen saturation, and mental status (Mandell, Wunderink et al. 2007) and (b) showing improvement from baseline on at least one symptom of cough, dyspnea, pleuritic chest pain, or sputum production, while not worsening on any of these four symptoms.
- In this analysis, improvement at Day 4 of symptoms along with stabilization of signs over the previous 24 hours was thought to be a reasonable choice of time for assessing the endpoint. But, Day 3 or 5 could perhaps also be used pending further analysis.
- The use of the early endpoint presumes that there is a later secondary outcome measure that captures overall outcome; relevant measurements such as temperature, respiratory rate, blood pressure, and oxygenation should be approached as supportive secondary measurements, and the IDSA/ATS guidelines provide a good reference for clinical stability based on vital sign measurements.

2.1.6 Data Not Yet Available and Needed for Project Team's Final Recommendations

(Please also refer to Section 2.2)

As summarized below, these analyses demonstrate the potential value of an early endpoint measure that is based on the symptoms of cough, chest pain, dyspnea, and sputum production. As demonstrated by the analysis of the ceftaroline registrational dataset, resolution of these symptoms in combination with demonstration of physiological stability (the temperature, heart rate, respiratory rate, blood pressure, oxygen saturation, and mental status stability parameters discussed above) offers an endpoint that offers advantages of a strong link to historical evidence of a substantial antibiotic treatment effect size relative to placebo and an objective approach to documenting improvement of the patient symptoms.

Although there are gaps in our knowledge regarding such an endpoint, the consensus opinion of the Project Team is that an endpoint based on these ideas could be used now to enable trials to proceed in this area. Additional work is needed to refine our understanding of such an endpoint, but there is a critical need for a bridge period with the use of interim efficacy endpoints. Thus, the ideas in this document are recommended for immediate use.

For the future, however, areas that require further clarification are

- Specific enrollment criteria
- Identification of alternative endpoints, including those that might be suitable for assessing response in patients with greater or lesser degrees of baseline severity of illness and symptoms. For example, critically ill patients may not be able to provide direct reporting on their symptoms.
- Are symptoms other than the four identified from these data relevant? Can the simple scoring scheme of Absent, Mild, Moderate, and Severe be better defined or made more robust?
- An approach to the important measures of clinical stability based on the temperature, heart rate, respiratory rate, blood pressure, oxygen saturation, and mental status stability parameters discussed above needs to be developed. Principally composed of physiological biomarkers, the FDA's approach to the ceftaroline dataset evaluated these relevant measures as elements of a composite outcome measure in that therapy was required to demonstrate an effect on symptoms as well as these related measures. Is this the most informative approach? The data that group evaluated showed that this would lower overall success rates as measured in current trials.
- Selection of the optimum time(s) for endpoint evaluation.
- Are alternative endpoint rules needed for drugs of other classes? The Project Team recognized that the data were derived based on drugs from a limited number of classes (beta-lactams, fluoroquinolones, and tetracyclines), that the pace of response might vary among drug classes, and that the endpoint rule might need to be reconsidered in the future as additional data for other drug classes become available.

2.1.7 Conclusions

1) There is strong support that an early clinical endpoint (e.g. Day 4, see below) of symptom improvement gives relevant data on how a patient feels and functions and provides evidence of a strong treatment effect size for antibiotics via its link to less well-defined assessments of symptom improvement in historical studies.

2) The four symptoms identified in the review to date (cough, pleuritic chest pain, dyspnea, and sputum production scored as Absent, Mild, Moderate, and Severe) are biologically relevant to the disease and are recommended. It may be possible to utilize other symptoms, but including others would require a new definition for what is considered a success and new datasets for analysis. Evaluations of whether all relevant symptoms are included in current definitions should be a focus for future research.

3) The overall measure proposed at present by the Project Team builds on these three elements:

   a) A one-point improvement in at least two symptoms and

   b) No worsening of any other symptoms with

   c) The assessment made on study Day[3] 4.

4) Assessment at Days 3 and/or 5 is also plausible, but measures at these times were more often discordant with overall clinical response in the available dataset. This finding is not robust as the differences may have been in part due to different numbers of observations on each Day. The extent of discordance is also dependent upon the response definition. Thus, Day 4 should be viewed as reasonable choice but also one that could be challenged by future data.

5) Of note, the proposed early clinical endpoint does not consider other interim events. Subjects who die before the Day 4 endpoint would lack data showing improvement and would of course be judged as Non-Responders. However, subjects who required a change in therapy due to a complication or adverse event might be judged at the early response timepoint as a Responder if initiation of alternative therapy produced an adequate response by Day 4. Although one might expect someone who received alternative therapy to be scored as a Non-Responder, the Project Team proposes scoring the early response measure based solely on clinical response. As the Project Team expects such discordant situations to be uncommon, the numerical impact on the early response endpoint should be insignificant. This type of event should be identified in secondary analyses.

6) There are important alternative viewpoints on the use of the proposed endpoint. In brief, the concerns focus on the limited data to support the new endpoint, the early endpoint's inability to capture the entire treatment course, and the potential challenge of using this endpoint in

[3] Throughout this document, Study Day 1 corresponds to the day of initiation of study therapy. An observation on Study Day 2 (usually the next calendar day) would be taken approximately 24h after initiation of therapy, an observation on Study Day 3 would be taken approximately 48h after therapy initiation, Study Day 4 would be approximately 72h after therapy initiation, and Study Day 5 would be approximately 120h after therapy initiation. The datasets discussed in this paper did not rigidly define specific time windows but rather appear to have followed a largely calendar-day based convention.

parallel with other endpoints as part of a global development program. These are discussed in detail in Section 3.2.

2.2 Phase 2: Qualitative Research Phase

The review of the available data by the Project Team revealed several research gaps in both defining all the relevant symptoms of importance to patients and in evaluating the reliability of measurements of patient symptoms. While it is critical to develop interim recommendations to allow drug development to proceed, it is equally critical to perform research to evaluate the validity and reliability of these recommendations or to improve upon them if needed. This research should be performed in as a timely a fashion as possible. It is planned that one or more research firms will be selected through a formal RFP process to complete a qualitative research phase of instrument development that would be based on both literature searches and patient interviews. This work might lead to improved outcome measures for future clinical trials in CABP.

The proposed studies will be conducted by a group of researchers highly experienced in the field of infectious disease, and will be guided by a Project Team that includes academic clinicians, drug development personnel from pharmaceutical companies, and representatives from the NIH, and the FDA.

Results from the retrospective clinical trial analyses and qualitative research studies will be used as input to designing prospective clinical studies to be conducted as part of a potential Phase 3, which would be proposed as a separate Biomarkers Consortium project and be focused on the design and conduct of one or more clinical studies to further test and validate specific endpoints and measurement approaches. While a standalone study cannot be ruled out, it is expected that these later studies will be able to be coordinated as companion studies to current trials being conducted by NIAID (National Institutes of Allergy and Infectious Diseases) or industry.

3 Interim Recommendations

3.1 Description of an Early Endpoint

1) Study design
    a) Most studies comparing one active agent with another would be of an non-inferiority design due to ethical and feasibility issues.
    b) Superiority trials are difficult to implement for serious or life-threatening infections unless there are no other active agents available. The one exception is add-on studies in which a second active agent is added to the base regimen, but achieving a superior effect over a fully dosed and active base regimen would be unlikely in setting where there is already effective therapy.
    c) Dose-response and placebo-controlled superiority study designs could be used in selected mild infections. Specific situations such as randomized dose-response trials and combination therapy trials do offer the tantalizing possibility of providing data on which to base the design of future non-inferiority trials.

d) However, an additional limitation is that the subjects who can be enrolled may have such limited and mild infection that the results cannot be generalized beyond the context of use in the given clinical trial to other patient groups with more severe forms of the illness.
e) Note that novel well-defined, reliable, and clinically meaningful endpoints can be used in superiority trials since there is no requirement for evidence of treatment effect from prior studies to evaluate assay sensitivity in the setting of superiority trials.

2) Endpoints
   a) Early assessment at Study Day 4, approximately 72h[4] after baseline measurement at time of randomization and treatment initiation, supports treatment effect by demonstration of
      i) A one-point improvement in at least two symptoms and
      ii) No worsening of any other symptoms
      iii) Where symptoms are Cough, Dyspnea, Pleuritic Chest Pain, and Sputum Production
      iv) And symptoms are scored as Absent (or none), Mild, Moderate, and or Severe.
   b) Later assessment at a fixed time point after initiation of therapy
      i) The Project Team did not debate the precise requirements for a later assessment endpoint and identified this as a topic for future research. Typical elements from prior studies would include
         (1) Survival,
         (2) Improvement (or resolution) of the clinical signs that are part of the early assessment endpoint,
         (3) Lack of a requirement for modification of therapy, and
         (4) Lack of adverse events leading to discontinuation of therapy.
      ii) The late assessment might or might not include a requirement to have been judged a Responder at the early endpoint (see the discussion on Alternative Viewpoints (Section 3.2).
      iii) To address the need for international harmonization of clinical trial design, the late endpoint could in fact be two time points; one at the end of therapy (EOT) and the other at an off-therapy (i.e., TOC) time point.
      iv) The best time(s) for the late endpoint(s) should be determined depending on the maximum length of treatment, the pharmacokinetic (PK)/pharmacodynamic characteristics of the drug, and the characteristics of the comparator agent.
      v) Assessments should be made at a fixed time point relative to the baseline measurement and study initiation that is the same across patients.
      vi) Collection of sufficient PK data to estimate individual subject drug exposure would allow for more complete exposure-response analyses for both early and late endpoints.
   c) Absence of elevated body temperature is not recommended as part of the early endpoint since it may be confounded by antipyretic therapy. Although persistent fever is occasionally due to a non-infectious cause such as drug-related fever, overall successful response without resolution of elevated body temperature would be unusual and its

[4] Throughout this document, Study Day 1 corresponds to the day of initiation of study therapy. An observation on Study Day 2 (usually the next calendar day) would be taken approximately 24h after initiation of therapy, an observation on Study Day 3 would be taken approximately 48h after therapy initiation, Study Day 4 would be approximately 72h after therapy initiation, and Study Day 5 would be approximately 120h after therapy initiation. The datasets discussed in this paper did not rigidly define specific time windows but rather appear to have followed a largely calendar-day based convention.

resolution is of interest to patients and physicians. It should thus be included as a sensitivity analysis and/or as part of a late assessment endpoint.

d) Parallel with the just-discussed issue of the resolution of elevated body temperature, improvement in the important measures of physiological clinical stability (e.g., the parameters suggested by the IDSA/ATS guidelines (Mandell, Wunderink et al. 2007)) would be expected but is not specifically part of the symptom-based endpoint described in this work. A conclusion of response based on symptoms without simultaneous achievement of such clinical stability would be unusual and would suggest an inter-current second process.

3) Study enrollment criteria
   a) This issue was outside of the scope of this project and was not discussed in detail by the Project Team. Diagnostic criteria similar to those in the March 2009 FDA Draft CABP guidance (Food and Drug Administration 2009, March) were presumed during Project Team discussions with key elements of standard clinical symptoms and PORT Risk Class of III or more. The issue of exclusion due to prior receipt of effective antibiotics was not analyzed by the Project Team. Likewise, the sample size challenge created by limiting the primary analysis to the microbiologically proven subset of patients was not discussed by the Project Team.
   b) As the proposed response endpoint rule requires improvement of at least one point for two symptoms, a minimum of two symptoms are required for study entry.

4) Although outside of the scope of this project and not discussed in detail by the Project Team, it was noted that late response should be assessed at fixed time points post-randomization or initiation of therapy to ensure a consistent duration of assessment time for successes and failures.

5) Proposed non-inferiority margin if applicable: This topic was not specifically discussed by the Project Team.

6) Sample size considerations: This topic was not specifically discussed by the Project Team.

7) Opportunities for harmonization globally
   a) See discussion above regarding choice of primary endpoint. These data could be presented to regulatory authorities in other countries for their evaluation. FDA members of the review group have offered to share these analyses with other regulatory agencies

8) Studies/ data needed to advance to final recommendations and timeframe for accomplishing same: Phase 2 data as described above.

3.2 Alternative Viewpoints, Issues, Limitations, and Areas for Future Work

There are alternative viewpoints within the Project Team regarding the conclusion that the primary measure should be taken at Day 4 of therapy. Although there was agreement among team members that the early measurement provided important information, some concerns were raised and should also be addressed in future research:

a) ***These endpoints rely in part on data from a very different medical era.*** Although biologically plausible, the specific proposal developed for elements of the proposed early endpoint is based on a small number of datasets, some of which are very old and which represent medical experience during an era that provided very different levels of supportive care.

b) ***Currently available agents are active for life-threatening infections such as CABP.*** Although there are demonstrated instances of detection of differences in efficacy or safety among agents for life-threatening infections as well as instances of detection of ineffective agents that were not subsequently registered for the given indication (e.g., daptomycin for pneumonia), currently available agents approved using traditional late assessment TOC endpoints are suitable to use as comparators in future trials (Spellberg 2011). As discussed below, some justification for this is that traditional late assessment TOC endpoints have always implicitly included a requirement for an early response, albeit not necessarily in a formal manner.

c) ***Recent pharmacometric analyses show a correlation between drug exposure and TOC endpoints***. A recent presented observation (Ambrose 2011; European Medicines Agency 2011) is that pharmacometric exposure-response analyses demonstrate a correlation of drug exposure with traditional clinical and microbiological endpoints.
    i) Arguments in favor of the plausibility of these correlations include:
        (1) The demonstrated relationships indicate that contemporary clinical endpoints (e.g. success or failure at the TOC) capture a measure of drug effect.
        (2) These analyses produce estimates of treatment effect relative to placebo which are similar to estimates derived from other sources but that are derived using current data from modern studies and thereby could negate concerns of meeting the constancy assumption.
        (3) The consistency of these observations (similar results can be shown for across both multiple indications [HAP, VAP, CAP, ABSSSI, and ABECB] and multiple antibiotic classes), the biological plausibility of the observations (drug effect should decline as exposure declines), the retention of the correlations when the analysis is controlled for age, severity of illness, or co-morbid disease, and the lack of an hypothesis regarding a host immune factor that would correspondingly alter drug exposure lend support to the need to consider carefully this approach.
        (4) In particular, this approach offers the possibility of validating non-inferiority margins using modern trial designs and modern endpoints. Moreover, pharmacometric exposure-response analyses offer the possibility of linking early and contemporary late clinical endpoints.

ii) This approach, however, can also be critiqued:
(1) Although these analyses are useful for identifying prognostic factors and generating hypotheses regarding plausible doses and schedules to be studied in properly conducted randomized trials, attempts at causal inferences from such analyses are biased due to confounding between treatment effects and prognostic patient characteristics.
(2) Specifically, it is not sufficient that exposure or organisms may be randomly assigned, since host factors are not randomly assigned and these latter factors cannot be adequately accounted for by matching. People with differing concentrations or minimum inhibitory concentrations can differ on other factors that affect outcome, like age, severity of illness, co-morbid disease, or many other covariates, and most of these are unidentified or unrecorded. Inherent differences in such patient characteristics are sufficiently influential to lead to substantial differences in concentrations; therefore it is likely that these inherent differences are also meaningfully predictive of the outcome measures.
(3) The consistency of results across settings may thus be explained by consistency of this same bias across those settings.

d) ***Early time points are already part of the traditional late assessment TOC outcome.*** An early measure of response is included in all clinical trials, but the timing and formality of this evaluation may differ from trial to trial and there is not a systematic requirement for investigators to make a final assessment at this time point. If improvement is not apparent at Day 3 or 4, the patient is generally withdrawn from study medication and the response defined as a failure for effectiveness analyses. These outcomes are carried forward for purposes of analyses at later time points. In some trials, this early assessment has been entirely informal and is captured only by noting whether the physician and patient continued the randomized therapy. In other trials, a formal recording a decision to continue has been taken. A systematic analysis of early time points with clear definitions of outcomes would help clarify the analysis of trial results. A great strength of the work presented here is that it provides a basis for documenting the reasoning that goes into the decision to continue or discontinue therapy at an early time point. Early and later time point assessments are not mutually exclusive and can both be measured in the setting of clinical trials.

e) ***Later endpoints provide a key overall perspective.*** While all team members agreed that early measurement provided important information on drug effects, some members of the Project Team believed that the primary outcome measure should be assessed at the EOT or beyond. The suggestion to use a later primary endpoint included these concerns:
i) Overall clinical cure at a late time point following EOT is important to evaluate durability of response and should be noted in the product labeling. Given that this measure thus takes on the role of being the principal measure that is relevant to the use of a drug, it could be argued that this measure best meets the ICH E9 (Section 2.2.2) test that: "The primary variable ('target' variable, primary endpoint) should be the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial."
ii) Use of the early endpoint as the primary study endpoint has not to date been specifically endorsed by other regulatory agencies. Global trial design harmonization is an important

goal and the full implications of the use of dual statistical analysis plans have yet to be understood by the community.

iii) As an example of a specific alternative approach, an overall endpoint which required both success at the early endpoint based on the rules proposed below (see Section 3) AND success at a later overall time such as at a typical TOC time point (preferably assessed using a direct measure of how a patient feels, functions or survives, lack of requirement for other therapy, lack of complications, etc.) could be considered to (i) incorporate the known effect size, (ii) capture the entire pattern of response, and (iii) address the concern that success rates inevitably rise over time such that even placebo-treated patients recover (or have died). If the effect size relative to placebo is sufficiently large for the early endpoint, the small number of patients who subsequently convert from success at the early endpoint to failure at the late endpoint (e.g., <5% in the tigecycline analysis) still supports a large effect size (**Figure 3**). This idea follows naturally from the critique of the traditional TOC endpoint discussed in Section 1.1 of this document and mimics the standard clinical (and clinical trial) practice of using a patient's early response to determine if therapy is adequate and suggests a connection between the demonstration of the ability of standard trial designs to detect inadequate drugs or exposures (see above). Combining the strength of a well-defined and objective early measure with the clinical relevance of the overall endpoint offers a potentially useful alternate option to the primary assessment at the time of the early endpoint. For example, such an approach might support international harmonization.

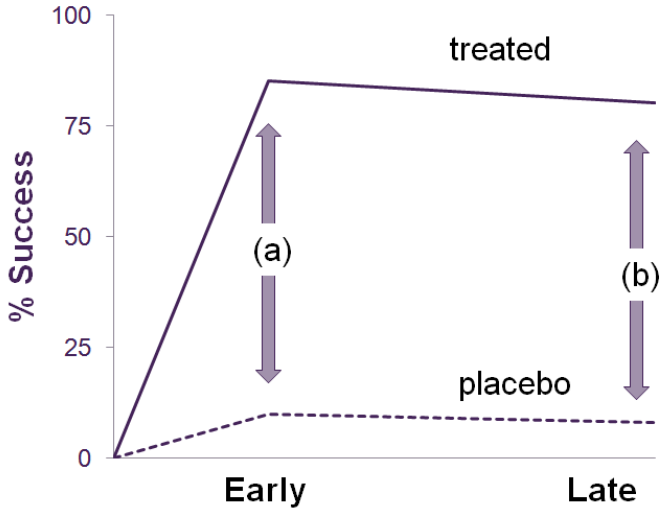


Figure 3. Estimating a late treatment effect using the estimate for an early treatment effect. If there is a large early effect AND if success at the late time (e.g., a typical TOC timepoint) requires early success, then a large treatment effect will still be present. As an example, the treatment effect at early times (a) for CABP is > 70% (see early sections of this document for data from Osler 1910, Bullowa 1937, Meakins 1939). In the data discussed by the working group, rates of discordance between the early endpoint and a late (TOC) clinical endpoint were < 5% (Section 2.1.3).

f) ***Time to response may provide useful insights.*** Since clinicians often choose to change therapy in the absences of response within two or three days, the early time point must have some value in addition to later time points evaluating durability of response or other variables such as relapse or adverse events. Time to response may also be an important measure but

not one for which at present there are data to pose a hypothesis for a non-inferiority trial. This is an approach that could be considered as additional data become available for analysis.

4 Conclusions

In the process outlined above various stakeholders including members from academia, industry and government agencies proposed interim, bridging outcome measures for registrational trials in the CABP indication.

These interim outcome measures are based on an evidence-based analysis of the historical literature that showed a treatment effect of antimicrobials in CABP based on symptom improvement at Day 4 after the first dose of study drug. While other outcome measures are relevant, there is insufficient evidence at present to base future non-inferiority trials solely on those outcomes. These outcomes could be studied by testing superiority hypotheses in future studies or possibly be based on new data such as the insights coming from recently presented pharmacometric exposure-response analyses.

The proposed early time point shows a substantial treatment effect for antimicrobials (approximately 30%; Section 2.1.2 above), allowing assessment of the non-inferiority of active agents at this time point. This large treatment effect (M1) provides a solid justification for selection of an M2 on the basis of clinical reasoning.

These interim outcome measures allows registrational studies to proceed while the Project Team plans future qualitative and quantitative research studies to evaluate the relationship between outcome measures in CABP and the operational characteristics of various measurement methods and time points in assessing outcomes in CABP. These future studies are critical in addressing knowledge gaps related to designing trials in CABP.

5 Supplemental Data

5.1 The Course of Untreated Pneumonia

1) The description provided by Osler in 1910 of the presentation of untreated pneumonia is particularly detailed (Osler 1910)
   a) When seen on the second or third day, the picture in typical pneumonia is more distinctive than any other acute disease.
   b) The patient lies flat in bed, often on the affected side; the face is flushed, particularly one or both cheeks; the breathing is hurried, accompanied often with a short expiratory grunt; the alae nasi dilate with each inspiration; … the eyes are bright; the expression anxious; and there is a frequent short cough which makes the patient wince and hold his side.
   c) The expectoration is blood-tinged and extremely tenacious.
   d) The temperature may be 104° or 105°.
   e) …
   f) After persisting for seven to ten days, the crisis occurs, and with a fall in the temperature the patient passes from the condition of extreme distress and anxiety to one of comparative comfort.

2) Osler provides these supplemental details in other parts of his review:
   a) Pain (pg. 174): "There is early a sharp, agonizing pain, generally referred to the region of the nipple or lower axilla on the affected side, and much aggravated on deep inspiration and coughing. It is absent in central pneumonia and much less frequent in apex pneumonia."
   b) Dyspnea (pg. 174): "Dyspnea is an almost constant feature. Even early in the disease the respirations may be 30 in the minute, and on the 2nd or 3rd day between 40 and 50. The movements are shallow, evidently restrained, and if the patient is asked to draw a deep breath he cries out with the pain."
   c) Cough (pg. 175): "This usually comes on with the pain in the side, and at first is dry, hard, without any expectoration. Later it becomes very characteristic – frequent, short, restrained, and associated with great pain in the side. In old persons, in drunkards, in the terminal pneumonias, and sometimes in young children, there may be no cough. After the crisis, the cough usually becomes much easier…"
   d) Sputum (pg. 174): "At first it may be mucoid, but usually after 24h it comes blood-tinged, viscid, and very tenacious. … in 100 cases in my clinic, in 16 there was little or no sputum, in 32 it was typically rusty, in 33 blood-streaked, in 3 cases very bloody. After the crisis the quantity is variable, abundant in some cases, absent in others"

3) Similar to Osler, Bullowa's 1937 description reinforces the sense of substantial morbidity but also gives insight into a steady deterioration during the early course of disease:
   a) "After four or five days, …
      i) …the patient who has become irritable and peevish, beings to "see things", is obstreperous, suspicious, and thinks he can take care of his own affairs. Under hypnotics, he may doze or become lethargic.

ii) By this time, the pain in his side has abated but the patient is distended and slumped in bed.
iii) He is cyanosed and breathes rapidly with effort.
iv) His pulse becomes rapid (120 or more), he refuses food and his weakness and emaciation are progressively severe
v) He becomes incontinent of stool and urine.

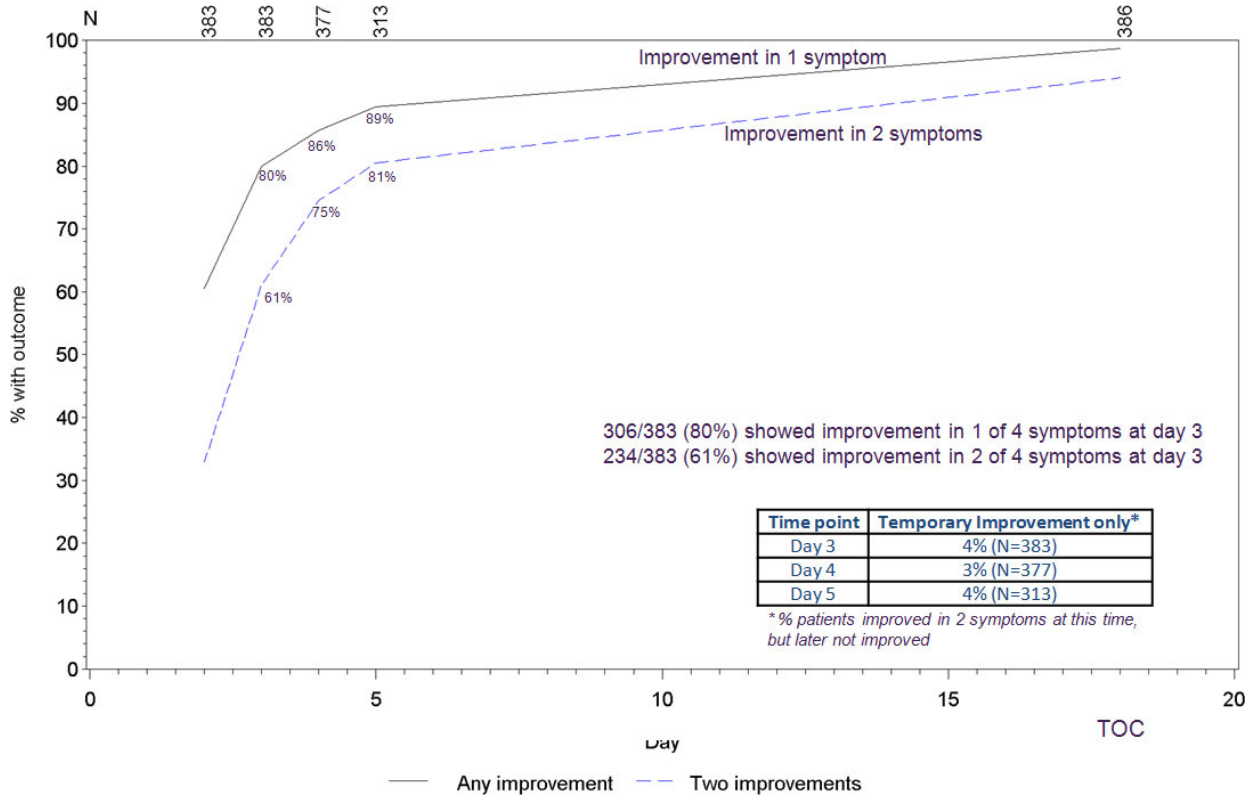
b) After eight or nine days, …
i) … the temperature falls following a drenching sweat. The patient then convalesces over several weeks, unless, after a few days there is an exacerbation of fever with the onset of a suppurative complication."

5.2 Supplemental Details from the Tigecycline-Levofloxacin CABP Dataset

Table 6: Frequency of baseline symptoms in the patient cohort

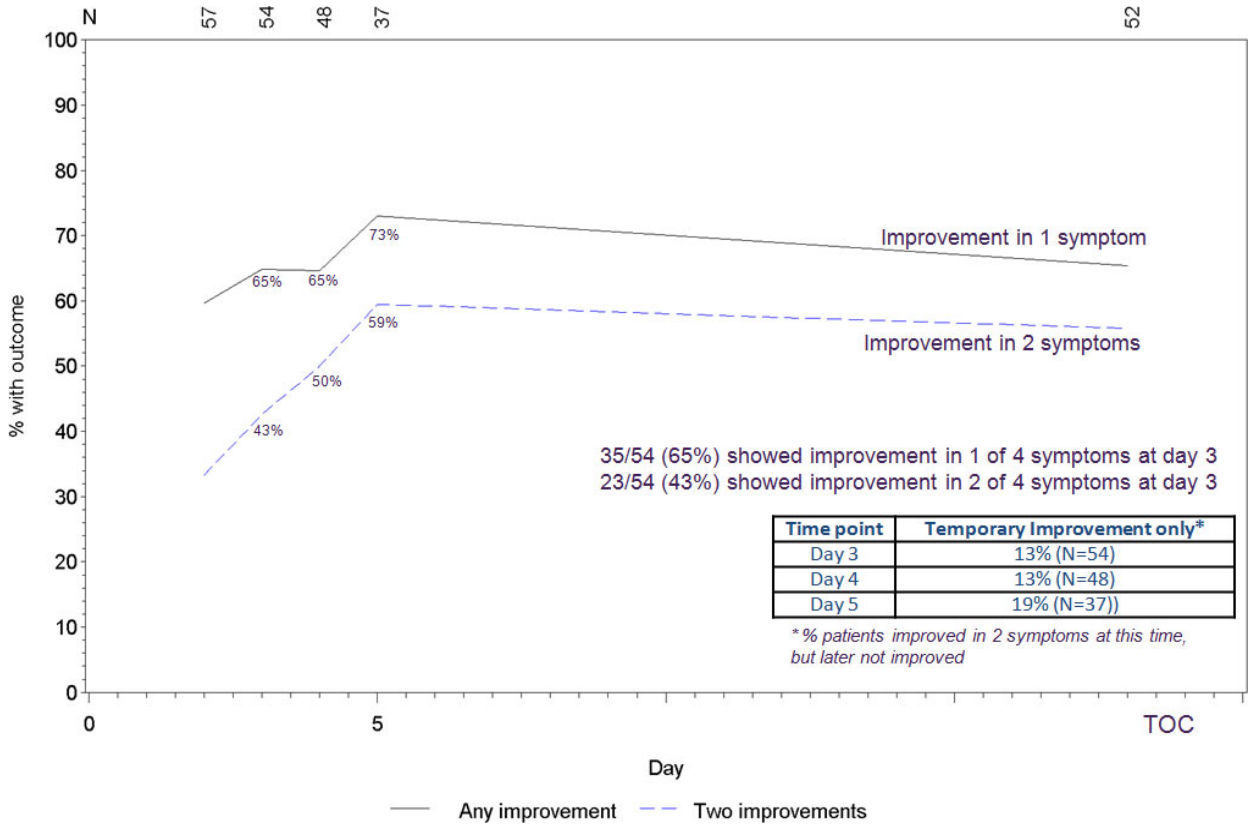
Number of symptoms for which the score at baseline was Mild, Moderate, or Severe	TOC clinical response		Total
	Cure	Failure	
1	15	4	19 (4%)
2	56	9	65 (14%)
3	136	21	157 (34%)
4	179	37	216 (47%)
Total	386	71	457

Figure 4. Improvement in CAP symptoms in patients judged Cured at the TOC visit



Temporary improvement rates use the total number of cures/total number of failures at TOC

Figure 5: Improvement in CAP symptoms in patients judged Failed at the TOC visit



Temporary improvement rates use the total number of cures/total number of failures at TOC

6 Project Team Members

The conclusions described within this document represent the work of the FNIH Biomarkers Consortium Project "Developing Endpoints for Clinical Trials of Drugs for Treatment of Acute Bacterial Skin and Skin Structure Infections and Community-Acquired Bacterial Pneumonia (Phases 1 and 2)".

- Joseph Toerner, M.D., M.P.H. (Co-chair; FDA, non-voting)
- George H. Talbot, M.D. (Co-chair; IDSA)
- Jeff Alder, Ph.D. (Bayer)
- Paul G. Ambrose, Pharm. D. (Institute for Clinical Pharmacodynamics; University of Oxford)
- Helen Boucher, M.D. (Tufts University)
- John Bradley, M.D. (University of California, San Diego)
- Laurie Burke, R.Ph, M.P.H. (FDA, non-voting)
- Aaron Dane, Ph.D. (AstraZeneca);
- Anita Das, Ph.D. (AxiStat)
- Dennis Dixon, Ph.D. (NIH/NIAID)
- Mike Dunne, M.D. (Durata Therapeutics)
- Barry Eisenstein, M.D. (Cubist)
- Thomas Fleming, Ph.D. (University of Washington)
- Dean Follmann, Ph.D. (NIH/NIAID)
- David Friedland, M.D. (Cerexa)
- Ian Friedland, M.D. (Cubist)
- Nickolas Kartsonis, M.D. (Merck)
- Achim Kaufhold, M.D., Ph.D. (Basilea)
- Scott Komo Dr.P.H. (FDA-observer, non-voting)
- Mike Kurilla, M.D., Ph.D. (NIH/NIAID)
- Kim Lindfield, Ph.D. (Cubist)
- Lily Llorens, Ph.D. (Cerexa)
- Susan Moriarty, M.D. (Cempra)
- Shawnmarie Mayrand-Chung, J.D., Ph.D. (NIH)
- Sumati Nambiar, M.D. (FDA-observer, non-voting)
- David Oldach, M.D. (Cempra)
- Gary J. Noel, M.D. (Paratek Pharmaceuticals)
- Elektra Papadopoulos, M.D. (FDA, non-voting)
- Roger Pomerantz, M.D. (Merck)
- John Powers, M.D. (SAIC in support of NIH/NIAID)
- Philippe Prokocimer, M.D. (Trius)
- John Quinn, M.D. (Pfizer)
- John H. Rex, M.D. (AstraZeneca)
- Jennifer Schranz, M.D. (Cempra)
- Judith A. Siuciak, Ph.D. (FNIH, non-voting)
- William Stubbings, Ph.D. (Basilea)
- David Wholley (FNIH, non-voting)

7 References

Ambrose, P. G. (2011). "Presentation on 21 April 2011 to the Project Team."

Boucher, H. W., G. H. Talbot, et al. (2009). "Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America." Clin Infect Dis **48**(1): 1-12.

Bullowa, J. G. M. (1937). Chapter II. The course, symptoms and physical findings. The management of pneumonias. New York, NY, Oxford University Press**:** 36-76.

el Moussaoui, R., B. C. Opmeer, et al. (2006). "Long-term symptom recovery and health-related quality of life in patients with mild-to-moderate-severe community-acquired pneumonia." Chest **130**: 1165-1172.

European Medicines Agency (2011). "Workshop on Antibacterials, Report of the Workshop, document EMA/257650/2011, available online at http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/04/WC500105473.pdf."

Finland, M., W. C. Spring, et al. (1940). "Immunological Studies on Patients with Pneumococcic Pneumonia Treated with Sulfapyridine." J Clin Invest **19**(1): 179-99.

Fleming, T. R. and J. H. Powers (2008). "Issues in Noninferiority Trials: The Evidence in Community- Acquired Pneumonia." Clin Infect Dis **47**: S108-S120.

Flippin, H. F., J. S. Lockwood, et al. (1939). "The treatment of pneumococcic pneumonia with sulfapyridine." JAMA-J Am Med Assn **112**: 529-534.

Food and Drug Administration, Center for Drug Evaluation, (2009, March). "Guidance for industry: Community-acquired bacterial pneumonia: Developing drugs for treatment." US Dept HHS, FDA, CDER **Available online at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm**.

Food and Drug Administration, Center for Drug Evaluation, (2010). "Guidance for industry: Non-inferiority clinical trials." US Dept HHS, FDA, CDER and CBER **Available online at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm**.

Mandell, L. A., R. G. Wunderink, et al. (2007). "Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults." Clin Infect Dis **44 Suppl 2**: S27-72.

Meakins, J. C. and F. R. Hanson (1939). "The treatment of pneumococcic pneumonia with sulfapyridine." Can Med Assoc J **40**: 333–6.

Metlay, J. P., S. J. Atlas, et al. (1998). "Time course of symptom resolution in patients with community-acquird pneumonia." Resp Med **92**: 1137-42.

Metlay, J. P., M. J. Fine, et al. (1997). "Measuring symptomatic and functional recovery in patients with community-acquired pneumonia." J Gen Intern Med **12**(7): 423-30.

Micheel, C., J. Ball, et al., Eds. (2010). Evaluation of biomarkers and surrogate endpoints in chronic disease. Washington, DC, National Academies Press.

Ochoa-Gondar, O., A. Vila-Corcoles, et al. (2008). "The burden of community-acquired pneumonia in the elderly: the Spanish EVAN-65 study." BMC Public Health **8**: 222.

Osler, W. (1910). Specific infectious diseases: Lobar pneumonia. The Principles and Practice of Medicine. New York, D. Appleton and Company**:** 164-192.

Pertel, P. E., P. Bernardo, et al. (2008). "Effects of prior effective therapy on the efficacy of daptomycin and ceftriaxone for the treatment of community-acquired pneumonia." Clin Infect Dis **46**(8): 1142-1151.

Petersdorf, R. G., L. E. Cluff, et al. (1957). "Pneumoccocal pneumonia treated with penicillin and aspirin." Bull Johns Hopkins Hosp **101**(1): 1-12.

Spellberg, B. (2011). "The tribulations of trials for antibacterial drugs." Clin Invest **(in press)**.

Spellberg, B., T. R. Fleming, et al. (2008). "Executive Summary: Workshop on Issues in the Design and Conduct of Clinical Trials of Antibacterial Drugs in the Treatment of Community-Acquired Pneumonia." Clin Infect Dis **47**: S105-S107.

Spellberg, B., G. H. Talbot, et al. (2008). "Position Paper: Recommended Design Features of Future Clinical Trials of Antibacterial Agents for Community-Acquired Pneumonia." Clin Infect Dis **47**: S249-S265.

Tanaseanu, C., C. Bergallo, et al. (2008). "Integrated results of 2 phase 3 studies comparing tigecycline and levofloxacin in community-acquired pneumonia." Diagn Microbiol Infect Dis **61**(3): 329-38.

Tanaseanu, C., S. Milutinovic, et al. (2009). "Efficacy and safety of tigecycline versus levofloxacin for community-acquired pneumonia." BMC Pulm Med **9**: 44.

Wilson, A. T., H. A. Spreen, et al. (1939). "Sulfapyridine in the Treatment of Pneumonia in Infancy and Childhood." JAMA **112**: 1435-1439.