Published in final edited form as:
Stat Biopharm Res. 2015 ; 7(4): 331–336. doi:10.1080/19466315.2015.1094406.

Comment: Fundamentals and Innovation in Antibiotic Trials

Scott R. Evans[1] and **Dean Follmann**[2]

[1]Harvard University

[2]National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH)

In this issue of *Statistics in Biopharmaceutical Research* (SBR), Price, Rubin, and Valappil discuss challenges associated with antibiotic development, outline issues in antibiotic trials, and discuss opportunities for innovative strategies. We commend the authors for their thorough review in this important medical area. We thank the Editors of SBR for the opportunity to comment on these challenging issues. In our commentary, we discuss some important points regarding trial fundamentals, aspects of trial design, and innovation in antibiotic trials.

Clinical Trial Fundamentals

The integrity and clinical utility of antibiotic trials can be improved with better application of clinical trial fundamentals. For example, the primary endpoint in many trials is measured at a test-of-cure (TOC) visit scheduled at a fixed time (e.g., 2 weeks) after the end-of-therapy (EOT). However the EOT can vary across patients depending on individual responses, implying that the timing of the TOC also varies. Scheduling the timing of the TOC in this manner is a clear violation of the standardization necessary to retain the benefits provided by randomization, e.g., between-arm differences observed at the end of the trial may be due to differences in treatment or differences in when responses were measured. The timing of the TOC should be anchored at a fixed interval after randomization. Another example is how indeterminate outcomes have become common to a degree of acceptance, again a violation of the ITT principle. Furthermore, surrogate endpoints (microbiological cure, i.e., clearance of bacteria) are sometimes used despite the availability of important clinical outcomes that directly measure how a patient feels, functions, or survives.

Noninferiority

The challenges with noninferiority (NI) trials are well-documented (Snappin, 2000; Powers, 2005; Fleming, 2008; Powers, 2008; Evans, 2009; Evans 2010; Hamasaki and Evans, 2013). These include the necessary assumptions of assay sensitivity and constancy, as well as the selections of the active control and NI margin.

The integrity of a NI trial can be threatened by poor assay sensitivity, i.e., a dilution of treatment differences through subtle design and conduct choices e.g., entry criteria, endpoint selection/timing, poor adherence, or loss to follow-up. For example prior/recue/concomitant therapy, common in the treatment of infections, can reduce assay sensitivity. Use of all-cause mortality as an endpoint may reduce assay sensitivity (but conversely using cause-specific death creates challenges with competing risks). Blinding provides limited protection for assay sensitivity, since blinded investigators can skew results toward similarity by assigning similar response ratings for all participants.

The integrity of antibiotic NI trials also depends on the constancy assumption (i.e., the effect of the control antibiotic has remained constant since the time it was proven effective). But constancy may be threatened by improvements in supportive medical care or—particularly relevant for antibiotics—the development of antibiotic resistance. If constancy does not hold then one could demonstrate NI, but both drugs could be ineffective.

NI is not transitive. A being noninferior to B, and B being noninferior to C, does not imply that A is noninferior to C. Thus when a new antibiotic is shown to be noninferior to a standard antibiotic in one trial, is itself selected as the active control for the next generation of antibiotic NI trials, there is a risk that the newest antibiotic is deemed noninferior and put into clinical use, but is truly inferior to the original control antibiotic (a process known as biocreep). Many recent antibiotic approvals were based on NI trials. From 2002-2009, 43 new medical entity approval packages were submitted to the FDA with about half for antimicrobials (United States Government Accountability Office, 2010). If these newly-approved antibiotics are utilized as controls in future NI trials, then biocreep could be a significant concern. For this reason, the active control should be the most effective agent under the conditions of the proposed trial (D'Agostino et. al., 2003).

NI margins should be selected to ensure that conclusions of NI imply: (1) new therapies are effective compared to placebo; and (2) clinically important levels of inferiority to control interventions can be ruled out, implying therapeutic exchangeability (Gau and Ware, 2008). Unfortunately, NI margin selection is often based on sample size considerations, or solely on criterion #1 with concerningly little attention paid to criterion #2 (ruling out clinically important differences).

A NI trial requires a margin which should be based on historical evidence of the magnitude of the active control compared to placebo for a specific endpoint. If there is no such evidence, a margin cannot be specified and a NI trial should not be conducted. Indeed, the validity of some completed trials that used clinical cure as an endpoint were questioned because it was realized there was no reliable evidence to justify a margin based on clinical cure, leaving the interpretation of the trial unclear (AIDAC 2012). Historical evidence is endpoint specific. Recent guidance on new endpoints for acute bacterial skin and skin structure infections used the change in lesion size because there were two studies conducted in Glasgow in the late 1930s that measured lesion size over time for patients treated with antibiotics compared to UV therapy (viewed as a proxy for a placebo) (FDA 2013). While newly developed endpoints based on patient reported outcomes are very attractive, it will be challenging to specify a margin for such endpoints.

An evaluation of a 50% random sample of the randomized controlled trials published in PubMed-indexed journals during 2011, concluded that 55 of 57 (96.5%) industry-funded NI trials had positive results (Flacco et. al., 2015). Another systematic review of 1992–2011 publications concluded that 325/337 (96.4%) of NI trials had positive results (Li et. al., 2013). This may be viewed as good news but it also raises concerns about the ability of NI trials to objectively evaluate therapies, or whether such results could be due to poor assay sensitivity, lack of constancy, ineffective control groups, and overly generous NI margins.

The FDA's Deputy Commissioner, Dr. Rob Califf has called for more pragmatic trials (Science, 2015). Pragmatic trials seek to inform clinical practice (e.g., will the antibiotic work in clinical practice?) rather than study disease biology (e.g., does the antibiotic kill bacteria?). Unfortunately the necessary requirements for reliable NI trials often limit pragmatism. Thus historically antibiotic trials have not been very pragmatic, and unfortunately often little effort is made to link the design and analysis of trials to practical decision-making in the clinical setting and manner in which antibiotics are utilized. For example, primary analyses are often based upon a subgroup of patients (e.g., modified-intent-to-treat population (mITT) that excludes patients without targeted pathogens) to help preserve assay sensitivity. Such evaluations are not directly helpful to patients or clinicians since a patient's subgroup-status (i.e., mITT) is unknown until after treatment has already been initiated due to a lack of accurate rapid diagnostics for identifying subgroup status. Furthermore entry criteria are often strict (e.g., limited to patients without prior therapy), again to preserve assay sensitivity. But this greatly limits generalizability since prior therapy is commonly observed in practice.

Superiority trials are inferentially stronger than NI trials and should be conducted when possible. While challenging for acute infections, delayed-start superiority trials may be acceptable for chronic infections where some delay in antibiotic initiation can be tolerated. A recent study randomized patients with extensively drug-resistant tuberculosis to either immediate initiation of linezolid or linezolid initiation two months later. The randomized interventions were in addition to their current therapy. The trial demonstrated a significant advantage of immediate linezolid versus delay (Lee et al 2012). While some settings could only allow short delays in therapy, even a small delay could allow for a superiority design with a sufficiently larger sample size.

External Controls

External-controlled trials (ECTs) have several potential advantages including: fewer prospectively enrolled participants, faster trials results, cost and resource efficiency, and attractiveness for trial participants since they have knowledge of treatment assignment. The major drawback of ECTs is that they are non-randomized studies. Randomization is the foundation for reliable statistical inference, ensuring the expectation of between-arm balance with respect to all factors, known or unknown, measured or unmeasured.

Because they are non-randomized studies, ECTs are vulnerable to all of the biases of observational studies. Bias can occur if the controls systematically differ from the prospective test group with regard to important factors in a manner that can affect outcome (e.g., supportive care, concomitant therapies, follow-up strategies, evaluation methods, or if patients with more favorable prognosis are selected for the prospective component of the trial but are included in the external component). Furthermore ECTs are not blinded and thus are subject to bias when eligibility or outcomes are assessed by clinicians or patients. In

ECTs, clinicians may also selectively prescribe additional therapies given knowledge of the treatment assignment.

ECTs should only be conducted when certain conditions hold (Gehan, 1984; ICH E10, 2000) although “even in such cases, however, there are documented examples of erroneous conclusions arising from such trials” (ICH E10). One of the requirements is that the external studies produce consistent, reliable results and the disease has a highly predictable course. However studies in the resistant pathogen setting display inconsistent results on objective outcomes such as all-cause mortality. A meta-analysis of deaths attributable to carbapenem-resistant Enterobacteriaceae infections (Falagas et al. 2014) showed substantial variation in mortality outcomes, with point estimates for survival ranging from 6% to 70% across nine studies. Research suggests that ECTs tend to produce “positive” results more frequently than randomized clinical trials (RCTs). Sacks et.al. (1983) reviewed 50 RCTs and 56 ECTs evaluating the same six therapies and found 76% of the ECTs but only 20% of the RCTs demonstrated superiority of the test group to the control. Given the heterogeneity of study results, the interpretation of ECT results would be very difficult to validate and interpret.

If the ECT is historical, then bias may occur due to factors that have changed since the time the historical control group data were collected (e.g., improvement in medical practice and patient standard of care, diagnostic criteria, or referral patterns). In a 10-year longitudinal study conducted at a single ICU (Rosenburger et. al., 2012), the mortality rate decreased despite the rise of resistant bacterial infections. The authors attributed the decrease in mortality to improvements in critical care and technology.

With evolving antibiotic resistance and changes in supportive care, the reliability of historical data is questionable. Therapies that worked yesterday, may not work well today or tomorrow. Trials conducted in the future will be conducted in diseases caused by organisms with a different resistance profile. Thus historical controls and test group participants will have different distributions of important baseline factors, challenging the reliability of historical controlled trials or Bayesian methods that use these data to support priors.

Given these concerns, randomization remains essential for reliable conclusions. The recent Ebola outbreak in West Africa was a crisis setting where some felt randomized trials were not necessary. Nonetheless, we would argue that a crisis does not change fundamentals and that future patients deserve proven therapies (Cox, Borio and Temple 2014). An RCT is currently being conducted in West Africa to see if Zmapp, a monoclonal antibody cocktail, improves survival over optimal supportive care (Dodd et al 2015).

Composite Endpoints

The advantages of the composite endpoint approach are numerous. First it may allow for superiority trial designs thus avoiding the complexities of NI trials described earlier, and potentially reduce sample sizes allowing for less costly and more feasible trials. The medical community has also called for more systematic evaluation of benefits and risk (or harms). A composite that incorporates benefits and harms directly addresses this challenge and has several advantages compared to other methods of benefit:risk assessment. Benefit:risk composites incorporate the association between benefits and harms, necessary for understanding the overall effects on individual patients. When applying an intervention, some patients may benefit while some may experience harms. If the patients experiencing harm and the patients experiencing benefit are largely disjointed, then it is important to identify ways to distinguish between these 2 groups to guide treatment selection. However, if the 2 groups are largely overlapping, then an assessment of the net effect (i.e., whether the benefits outweigh the harms) is needed. The traditional approach of separate analysis of each endpoint cannot distinguish between these 2 scenarios and thus does not optimally evaluate the distribution of the totality of the effects on individual patients. Analysis of the composite is more pragmatic in the sense it provides direct information regarding the overall patient outcomes, resulting in greater utility for patients and clinicians. Furthermore incorporating competing risks as part of the composite, alleviates the challenges of trying to interpret the results on individual outcomes that are affected by competing risks (Evans et.al., 2015).

Given these advantages composite endpoints that combine benefits and harms (and possibly quality-of-life (QoL) outcomes) are a particularly promising area. The Antibacterial Resistance Leadership Group (ARLG) is using Response adjusted for duration of antibiotic risk (RADAR), a novel methodology utilizing a superiority design and a 2-step process: (1) categorizing patients into an overall clinical outcome (based on benefits and harms), and (2) ranking patients with respect to a desirability of outcome ranking (DOOR). DOORs are constructed by assigning higher ranks to patients with (1) better overall clinical outcomes and (2) shorter durations of antibiotic use for patients similar overall clinical outcomes. DOOR distributions are compared between antibiotic use strategies. The probability that a randomly selected patient will have a better DOOR if assigned to the new strategy is estimated (Evans et.al., 2015). Others note that composite/DOOR could change the paradigm in antimicrobial stewardship (Molina and Cisneros, 2015).

The DOOR concept may have broader utility in antibiotic development. For example, colistin is an antibacterial drug that was discovered in the 1940s and largely abandoned for several decades, but has more recently undergone revived use due to activity against several Gram-negative pathogens that cause life-threatening infections and are resistant to multiple other antibiotics (Flagas and Kasiakou, 2005). However, colistin has questionable efficacy (Paul et.al. 2010) and causes nephrotoxicity and neurotoxicity (Koch-Weser et.al. 1970; Wolinsky and Hines, 1962; Hartzell et.al., 2009). A new drug could provide a superior alternative to colistin if it either: (i) improves efficacy on major outcomes such as mortality, or (ii) has similar efficacy, but reduces rates of clinically meaningful adverse effects. In a randomized trial comparing colistin to a new therapy, a DOOR could be defined based on an ordinal composite clinical outcome as follows:

- Survives without a major adverse event
- Survives with a major adverse event
- Death

If the new drug reduces major adverse events, then a trial comparing trial participants using DOOR may have greater power than a mortality trial to detect a benefit over colistin. It would be important to utilize major adverse events of unquestioned importance to the patient (e.g., irreversible renal failure or the need for hemodialysis). If the adverse events were loosely defined (e.g., reversible creatinine clearance changes that would go unnoticed by trial participants), then reductions in less meaningful adverse events could falsely imply overall benefit, such as when the new drug increases mortality relative to colistin, or when neither drug reduces mortality relative to a placebo.

One disadvantage is that construction of the composite is often novel and challenging. Careful deliberation is needed to synthesize the outcomes typically measured in trials of a particular clinical disease. The ARLG is conducting a pre-trial sub-study to develop and validate use of a composite DOOR strategy in a future *Staphylococcus aureus* bacteremia trial. Twenty representative patient profiles (including benefits, harms, and QoL) were constructed based on experiences observed in completed trials in *Staphylococcus aureus* bacteremia. The profiles are being independently ranked by each member of a group of expert clinicians. The correlations among the ranks are being examined to evaluate consensus. An average rank is being generated for each profile. Outcome characteristics that guide the average ranks will be evaluated to develop an objective algorithm for objectively ranking patient outcomes. Future trials could utilize the ranking strategy by comparing the distribution of ranks between randomized therapies.

Another disadvantage of composite approach is that the components that comprise the composite could have differing levels of importance, and the result on the composite could be primarily driven by components of lesser importance, effectively not providing appropriate weight to the most important component(s). Furthermore, significance on the composite does not imply significance on the components, and significance on the components does not imply significance on the composite. Thus a fundamental part of composite endpoint analyses is a careful evaluation of the effects on each component (Neaton et.al., 2005). A strategy that could be employed to address this concern is to specify and evaluate co-primary endpoints (Sozu et.al., 2015), the composite and the most important component, to ensure that effects on the most important component are not hidden by the composition.

Multiple Body Sites

While conceptually appealing, combining information across different body sites requires careful thought. One must specify a method of analysis for combination which implicitly assumes a degree of combinability across body sites. There would be little statistical power to assess if the assumed degree of combinability is correct, furthermore, it is dangerous to have a data-dependent choice of test. Thus the decision to combine across body sites would need to be based on data available when the study was designed and expert opinion. The authors describe a few approaches which we embellish and extend.

The methods can be ordered by the level of assumed similarity of effect across body sites. The strongest assumption would be to simply use the same endpoint e.g., clinical cure or mortality and ignore body site, effectively viewing body site as irrelevant. The hierarchical approach mentioned is an off-the-shelf technology that specifies a single model for all of the data. One example is

$$\text{logit}\left\{P\left(Y_{ij}=1\right)\right\}=\beta_0+\beta_1 Z_{ij}+b_j$$

where i=1,…,m_j indexes subjects at a body site j, j=1,…,B indexes body sites, Z_{ij} is the indicator for new drug for person i at body site j, Y_{ij} is the indicator of success for person i at body site j, and b_j s a random body site effect assumed to be normal with mean 0 and variance V. The parameter β_1 is the effect of new drug on the log-odds scale while $\beta_0 + b_j$ provides the effect of body site j. The variance V, which is estimated, describes the degree of similarity across body sites: if V is zero, the model treats body site as irrelevant.

In principle meta-analytic methods could also be used. Here an estimated treatment effect for a given body site would be calculated along with an estimate of its variance. The treatment effects could be combined using fixed or random effects meta-analysis. The former is similar to a stratified analysis where strata correspond to body sites.

Conclusions

In the late 1970s, in response to a sense of complacency about the threat of infectious disease following the mature miracle of antibiotics, Richard Krause, former Director of the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH), presciently spoke of the ”restless tide” of pathogens that would ceaselessly evolve and adapt to changing environments including those defined by antibiotics (Krause 1981). Today there is broad concern about the development of antibiotic resistance and a sense that new antibiotics are urgently needed to deal with the seemingly inevitable crisis. The statistical and regulatory challenge is to use and develop strategies and methods that result in reliably proven drugs being placed in the medical armamentarium. The NI paradigm, which dominates anti-infective drug development, is challenging in many ways that are not present with superiority studies, and careful design and execution of a NI trial is essential. Innovative methods are appealing, but need to be thoughtfully and responsibly evaluated to ensure that they do not effectively lower the bar for licensure in an opaque manner. Transparent and calculated changes to conventional approaches such as requiring one rigorous phase III study instead of two are worth considering. Evidence from the device and government-funded (e.g., NIH funded) setting demonstrate that a single trial strategy can work. Nonetheless, even in response to evolving pathogens, it is important that traditional principles such as randomization, blinding, appropriate endpoints, and strong control of error rates be maintained.

References

AIDAC. 2012. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Anti-InfectiveDrugsAdvisoryCommittee/UCM340465.pdf

Couzin-Frankel J. Clinical trials get practical. Science. 348:6233. Pg. 382

Cox E, Borio L, Temple R. Evaluating ebola therapies ---- The case. for RCTs New England Journal of Medicine . 2014:2350–2351. [PubMed: 25470568]

D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Stat Med. 2003; 22:169–86. [PubMed: 12520555]

Dodd LE, Proschan MA, Neuhaus J, Koopmeiners J, Neaton J, Beigel JD, Barrett K, Lane HC, Davey RT. Design of a randomized controlled-trial for Ebola virus disease medical countermeasures: the Ebola MCM study. Unpublished manuscript. 2015

Evans S. Noninferiority clinical trials. Chance. 2009; 22:53–8.

Evans SR. Estudos Clinicos de Nao-Inferioridade. Revista Brasileira de Medicina. 2010; 67:7.

Evans SR, Rubin D, Follmann D, Pennello G, Huskins WC, Powers JH, Schoenfeld D, Chuang-Stein C, Cosgrove SE, Fowler VG Jr, Lautenbach E, Chambers HF. Desirability of Outcome Ranking (DOOR) and Response Adjusted for Duration of Antibiotic Risk (RADAR). CID. 2015; 61(5):800–806.

FDA. Guidance for industry acute skin and skin structure infections: Developing drugs for treatment. 2013. http://www.fda.gov/downloads/Drugs/Guidances/ucm071185.pdf

Flagas ME, Kasiakou SK. Colistin: the revival of polymyxins for the management of multidrug-resistant Gram-negative bacterial infections. Clinical Infectious Diseases. 2005; 40:1333–1341. [PubMed: 15825037]

Falagas ME, Tansarli GS, Karageorgopoulos DE, Vardakas KZ. Deaths attributable to carbapenem-resistant Enterobacteriaceae infections. Emerging Infectious Diseases. 2014; 20(7)10.3201/eid2007.121004

Flacco ME, Manzolia L, Boccia S, Capasso L, Aleksovska K, Rosso A, Scaioli G, De Vitoe C, Siliquinif R, Villarie P, Ioannidis JPA. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. Journal of Clinical Epidemiology. 2015; 68:811–820. [PubMed: 25748073]

Fleming TR. Current issues in non-inferiority trials. Stat Med. 2008; 27:317–32. [PubMed: 17340597]

Gao P, Ware JH. Assessing non-inferiority: a combination approach. Stat Med. 2008; 27:392–406. [PubMed: 17575568]

Gehan EA. The evaluation of therapies: historical control studies. Statistics in Medicine. 1984; 3:315–324. [PubMed: 6528131]

Hamasaki T, Evans SR. Noninferiority Clinical Trials: Issues in Design, Monitoring, Analyses, and Reporting. Igaku no Ayumi. 2013; 244(13):1212–1216.

Hartzell JD, et al. Nephrotoxicity associated with intravenous colistin (colistimethate sodium) treatment at a tertiary care medical center. Clinical Infectious Diseases. 2009; 48:1724–1728. [PubMed: 19438394]

International Conference on Harmonization (ICH)-E10. Choice of Control Group and Related Issues in Clinical trials. 2000

Koch-Weser JK, et al. Adverse effects of sodium colistimethate: manifestations and specific reaction rates during 317 courses of therapy. Ann Intern Med. 1970; 72(6):857–868. [PubMed: 5448745]

Krause, RM. The restless tide: The persistent challenge of the microbial world. The National Foundation for Infectious Diseases; Washington DC; 1981.

Lee M, Lee J, Carroll MW, Choi H, Min S, Song T, Via L, Goldfeder L, Kang F, Jin B, Park H, Kwak H, Kim H, Jeon S, Jeong I, Joh J, Chen R, Olivier K, Shaw P, Follmann DA, Song S, Lee JK, Lee D, Kim C, Dartois V, Park S, CHo S, Barry C. Linezolid for the Treatment of Chronic Extensively Drug-Resistant Tuberculosis. New England Journal of Medicine. 2012; 367(16):1508–1518. [PubMed: 23075177]

Li Y, He Y, Sheng Y, et al. Systematic evaluation of non-inferiority and equivalence randomized trials of anti-infective drugs. Expert Rev AntiInfect Ther. 2013; 11:1377–89.

Molina J, Cisneros JM. A Chance to Change the Paradigm of Outcome Assessment of Antimicrobial Stewardship Programs. CID. 2015; 61(5):807–808.

Neaton J, et al. Key issues in end point selection for heart failure trials: Composite endpoints. J Card Fail. 2005; 11(8):567–575. [PubMed: 16230258]

Paul M, et al. Effectiveness and safety of colistin: prospective comparative cohort study. J Antimicrob Chemoth. 2010; 65:1019–1027.

Powers JH, Cooper CK, Lin D, Ross DB. Sample size and the ethics of non-inferiority trials. Lancet. 2005; 366:24–5. [PubMed: 15993221]

Powers JH. Noninferiority and equivalence trials: deciphering “similarity” of medical interventions. Stat Med. 2008; 27:343–52. [PubMed: 18186148]

Snappin SM. Noninferiority trials. Curr Control Trials Cardiovasc Med. 2000; 1:19–21. [PubMed: 11714400]

Sozu, T.; Sugimoto, T.; Hamasaki, T.; Evans, SR. Sample Size Determination in Clinical Trials with Multiple Endpoints . Springer Cham/Heidelberg; New York: 2015.

United States Government Accountability Office. Report to Congressional Requesters. New Drug Approval: FDA's Consideration of Evidence from Certain Clinical Trials. Jul. 2010 http://www.gao.gov/assets/310/308301.pdf

Wolinsky E, Hines JD. Neurotoxic and nephrotoxic effects of colistin in patients with renal disease. N Engl J Med. 1962; 266:759–762. [PubMed: 14008070]