







# Evaluating the desirability of outcome ranking and response adjusted for duration of antibiotic risk for clinical trials of antibiotics in pediatric pneumonia

John M. VanBuren<sup>\*1</sup> , Russell K. Banks<sup>1</sup>, Nathan Kuppermann<sup>2,3</sup> , Jeffrey S. Gerber<sup>4,5</sup> , Richard M. Ruddy<sup>6,7</sup> , T. Charles Casper<sup>1</sup> , Todd A. Florin<sup>8,9</sup> , Pediatric Emergency Care Applied Research Network (PECARN)

<sup>1</sup>Division of Critical Care, Department of Pediatrics, University of Utah, Salt Lake City, UT, United States

<sup>2</sup>Department of Emergency Medicine, George Washington University School of Medicine and Health Sciences, Washington, DC, United States

<sup>3</sup>Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, DC, United States

<sup>4</sup>Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, United States

<sup>5</sup>Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA, United States

<sup>6</sup>Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, United States

<sup>7</sup>Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

<sup>8</sup>Division of Emergency Medicine, Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, United States

<sup>9</sup>Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

\*Corresponding author: John M. VanBuren, 295 Chipeta Way, Salt Lake City, UT (john.vanburen@hsc.utah.edu)

## Abstract

Objective outcomes for pediatric community-acquired pneumonia (CAP) are lacking. The desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR) outcome encompass clinical benefit and adverse effects, while also accounting for antibiotic exposure. We evaluated DOOR and RADAR (DOOR/RADAR) through simulations and compared sample-size considerations to noninferiority designs in a hypothetical trial comparing antibiotics and no antibiotics (ie, placebo) for children with mild CAP. We also evaluated a trial comparing different durations of antibiotic therapy. Three scenarios were considered: 1 with no difference in DOOR between the 2 groups, 1 in which placebo is more efficacious, and another in which amoxicillin is more efficacious than placebo. The power to detect a difference between arms was greater using DOOR/RADAR compared with DOOR alone. Assuming a sample size of 200, DOOR had 2.5%, 50%, and 65% power to detect a statistical difference between arms for scenarios 1–3, respectively, significantly less than DOOR/RADAR. Importantly, DOOR/RADAR incorrectly identified placebo as superior in scenario 3, where amoxicillin was truly efficacious. Sample size requirements for noninferiority designs were larger to achieve similar levels of power as DOOR and DOOR/RADAR. DOOR/RADAR has the potential to lead to an incorrect conclusion declaring placebo superior when amoxicillin is efficacious.

**Key words:** DOOR; RADAR; pneumonia; pediatrics; noninferiority.

## Introduction

In the United States, the annual incidence of hospitalization for community-acquired pneumonia (CAP) in children is approximately 15.7 per 10 000,<sup>1</sup> making it the second most common reason for pediatric hospitalization.<sup>2</sup> Despite its prevalence and importance, substantial variability exists in the outcomes assessed in observational studies and clinical trials of pediatric CAP. Commonly used outcomes often are nonspecific, such as hospital length of stay or revisit rates, or occur in few children, such as sepsis.

Given these limitations, the need for objective outcome measures in CAP in children is highlighted as a critical area for future research by the Pediatric Infectious Diseases Society and Infectious Diseases Society of America.<sup>3</sup> This outcomes gap is a major barrier to harmonization in research and translation

into clinical practice for trials of antibiotics in pediatric CAP.<sup>4</sup> In addition to variability in outcomes, clinical trial design has been a challenge. Antibiotic trials have traditionally used noninferiority designs,<sup>5</sup> which do not answer the question of whether 1 approach is better than another.<sup>6</sup> In addition, noninferiority trials are prone to bias and manipulation due to their complexity and the choice of the noninferiority margin, in addition to requiring large and potentially infeasible sample sizes.<sup>7–11</sup>

To overcome the limitations of noninferiority trial designs, a novel outcome has been proposed: the desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR).<sup>6</sup> This outcome combines clinical benefit and adverse effects to examine the totality of impact on individuals, including benefits and harms. First, patients are assigned a DOOR, which is an ordinal ranking based on an overall clinical outcome

Received: October 11, 2023. Accepted: July 19, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

**Table 1.** Overall desirability of outcome ranking clinical outcome ranks, from most to least desirable.

| Rank | Clinical outcome  |   |
|------|---|---|
|      | Clinical response   | Side effects and serious adverse events |
| 1    | Adequate with symptom resolution  | None                                    |
| 2    | Adequate with symptom resolution  | Mild                                    |
| 3    | Adequate with symptom resolution  | Moderate                                |
| 4    | Adequate with symptom resolution  | Severe                                  |
| 5    | Adequate with persistent fever, work of breathing, tachypnea, cough                     | None or any grade                       |
| 6    | Inadequate with additional emergency department or clinic visit without hospitalization | None or any grade                       |
| 7    | Inadequate with hospitalization   | None or any grade                       |
| 8    | Death   | Not applicable                          |

that includes response to therapy and adverse effects. Second, patients with the same DOOR are further discriminated by antibiotic exposure (RADAR), with the premise that less antibiotic exposure is better, given the same clinical outcome. Those with better clinical outcome ranks will always receive a better overall rank regardless of antibiotic exposure. The DOOR and RADAR (hereafter, DOOR/RADAR) outcome allows clinical trials to assess superiority of 1 arm over another, as opposed to a noninferiority design traditionally used by many prior CAP antibiotic trials. Critics of DOOR/RADAR have outlined several concerns, including the selection and number of the DOOR ranks (subjective, not disease specific, and lack the patient perspective),<sup>12–14</sup> the potential for missing imbalances in distributions of the individual DOOR ranks,<sup>12,15</sup> lack of power to assess rare but severe outcomes,<sup>16,17</sup> challenges in interpretation of the effect size,<sup>13</sup> and the linkages between exposure and outcomes.<sup>17</sup>

Thus, we sought to evaluate the utility of DOOR/RADAR for different outcomes for clinical trials of antibiotics in children with mild CAP, which we define as a child with CAP well enough to be treated as an outpatient.<sup>3,18</sup> In addition, we compared, through statistical simulations, sample-size scenarios for DOOR/RADAR with noninferiority designs. Given current interest and the importance of optimizing antibiotic use in pediatric CAP, we performed these simulations in a hypothetical trial evaluating antibiotics vs no antibiotics (ie, placebo) for mild pediatric CAP, in addition to a trial evaluating different durations of antibiotic therapy. These results can be used to prepare for clinical trials evaluating different antibiotic treatment strategies in young children with mild CAP.

## Methods

### DOOR/RADAR

For DOOR/RADAR simulations, we used the DOOR categories presented in the Short vs. Standard Course Outpatient Therapy of Community-Acquired Pneumonia in Children (SCOUT-CAP) trial that compared a short course (5 days) of antibiotics with a “standard course” (10 days) in children with mild CAP (Table 1).<sup>6,18</sup> The 8 categories range from adequate clinical response with symptom resolution without adverse events to death. Categories 1–4 capture increasingly severe antibiotic side effects among participants who meet the definition of adequate clinical response with symptom resolution. Categories 5–8 represent increasing degrees of clinical severity, regardless of antibiotic side effects. Following the methods of Evans et al.,<sup>6</sup> we derived DOOR/RADAR by first ranking patients according to DOOR and then by considering the number of days of antibiotic administration (RADAR). When the clinical outcomes for 2 simulated patients were different, the patient with the better clinical outcome (ie, DOOR) received a higher rank,

regardless of antibiotic exposure. When the DOOR of 2 simulated patients was the same, the patient with shorter duration of antibiotic therapy received a higher rank. For instances where death occurred, RADAR was not considered, and patients were given the worst possible rank. For these simulations, the assumption was made that all patients did not receive antibiotics prior to trial enrollment, because we anticipated that prior antibiotic use would exclude a child from a randomized trial evaluating antibiotic use.

To understand the potential implications of using DOOR and/or DOOR/RADAR as outcomes in future trials, we evaluated 3 scenarios (Table 2). For these scenarios, we considered hypothetical clinical trials in which patients in 1 arm were prescribed antibiotics and those in the other arm were not (ie, they received a placebo). These trials evaluate whether a “no antibiotic” strategy is superior to antibiotics for mild CAP, using a similar framework to showing a short course is superior to standard course in the SCOUT-CAP trial.<sup>18</sup> The theoretical framework for this comparison is grounded in the premise that most cases of mild CAP in preschool-aged children likely has a viral cause.<sup>19</sup> This is consistent with the Infectious Diseases Society of America and Pediatric Infectious Diseases Society pediatric CAP guidelines that recommend against routine antibiotics in preschool-aged children well enough to be treated as outpatients.<sup>3</sup>

Assumed distributions of the DOOR categories (Table 2) were determined based on a combination of data from SCOUT-CAP, other trials of antibiotic use in mild pediatric CAP, and the authors’ expert opinions.<sup>18,20,21</sup> Scenario 1 represents the clinical case in which there is no difference in the distribution of patients by treatment arm among DOOR categories (the “null” scenario); that is, the overall status of the patient’s recovery is not influenced by treatment assignment. Scenario 2 represents a clinical case in which placebo is more efficacious than amoxicillin, due to fairly equivalent clinical outcomes but reduced antibiotic side effects in the placebo group (“efficacious placebo”). Scenario 3 is a hypothetical scenario used for illustration in which amoxicillin is substantially more efficacious with respect to clinical response than the placebo (“efficacious amoxicillin”).

For all scenarios, we estimated the distribution of antibiotic exposure using the maximum presumed number of days of antibiotic exposure and the probability of the model selecting the maximum number. Statistically, this would reflect the *N* and probability used in simulating antibiotic-days from a binomial distribution for each patient (Table 2). For example, in the amoxicillin arm of DOOR outcome 1 (first row of the Table 2), we assume there was a 90% chance a given patient took the antibiotic each day. In the placebo arm for this same outcome, there would be some patients who would end up taking antibiotics even though they were assigned to the placebo arm. Specifically, for this example,

**Table 2.** Scenario parameters evaluated through simulation.

| DOOR outcome <sup>a</sup> | Null, % (scenario 1) |         | Efficacious placebo, % (scenario 2) |         | Efficacious amoxicillin, % (scenario 3) |         | Antibiotic-days distribution, n (%) <sup>b</sup> (scenarios 1-3) |         |
|---------------------------|----------------------|---------|-------------------------------------|---------|---|---------|--|---------|
|                           | Amoxicillin          | Placebo | Amoxicillin                         | Placebo | Amoxicillin                             | Placebo | Amoxicillin  | Placebo |
| 1                         | 53                   | 53      | 53                                  | 69      | 53                                      | 46      | 7 (90)   | 7 (5)   |
| 2                         | 23                   | 23      | 23                                  | 14      | 23                                      | 14      | 7 (70)   | 7 (5)   |
| 3                         | 8                    | 8       | 8                                   | 2       | 8                                       | 3       | 7 (60)   | 7 (5)   |
| 4                         | 2                    | 2       | 2                                   | 1       | 2                                       | 1       | 7 (50)   | 7 (5)   |
| 5                         | 10                   | 10      | 10                                  | 10      | 10                                      | 13      | 7 (90)   | 7 (5)   |
| 6                         | 3                    | 3       | 3                                   | 3       | 3                                       | 10      | 10 (90)  | 10 (90) |
| 7                         | 0.95                 | 0.95    | 0.95                                | 0.95    | 0.95                                    | 8.95    | 15 (95)  | 15 (95) |
| 8                         | 0.05                 | 0.05    | 0.05                                | 0.05    | 0.05                                    | 4.05    |  |         |

<sup>a</sup>Desirability of outcome ranking (DOOR) and DOOR plus response adjusted for duration of antibiotic risk (RADAR) are derived by first ranking patients according to DOOR, then by the number of days of antibiotic administration. For example, in the amoxicillin arm of DOOR outcome 1, we assumed there is a 90% probability of taking antibiotics on any given day. In the placebo arm, there will be some patients who will end up taking antibiotics even though they were assigned to the placebo arm. Specifically, for this example, we assumed these patients who receive placebo will have a 5% probability of taking antibiotics on any given day. In this case, the statistical probability of taking 0 days of antibiotics is approximately 70%. The sampling distribution of antibiotic-days was applied uniformly within each DOOR outcome value and across all 3 scenarios. Scenarios are defined solely by changes to the sampling distribution of DOOR by study arm (amoxicillin vs placebo).

<sup>b</sup>Data are reported as n (%) of the binomial distribution.

we assumed these patients in the placebo group would have a 5% probability of taking the antibiotic each day. In this case, the statistical probability of taking 0 days of antibiotics was approximately 70%. Higher levels of DOOR categories are assumed to have likelihoods of taking antibiotics to reflect the clinical care anticipated if the participant returned to a health care facility or was hospitalized. Visually, these distributions are shown in Figure S1. Ranking patients by DOOR and then by the days of antibiotic administration defines the DOOR/RADAR outcome, but scenarios were solely defined by changes to the hypothetical sampling distributions of the DOOR outcome. In addition to the 3 scenarios presented here, we simulated a different null case in which there was no difference in DOOR/RADAR combined distribution to confirm the overall type I error rate in simulations. We simulated 3 additional scenarios assuming a 2-arm trial comparing a 7-day antibiotic course with a 3-day antibiotic course (Table S1).

For each set of scenarios and both DOOR and DOOR/RADAR outcomes, we performed 10 000 simulations under various sample-size scenarios and recorded whether the trial resulted in statistical significance, regardless of the direction of significance. In addition, we estimated the probability that the outcome of a participant randomized to the placebo arm was superior to that of a patient randomized to the amoxicillin arm, using previously published methods.<sup>18</sup> For each simulated trial, we quantified what the results would be if statistical gatekeeping were used, first evaluating the DOOR distribution, then the RADAR distribution if the probability of better DOOR for shorter antibiotic durations surpassed a threshold. Gatekeeping is a procedure whereby a specified hypothesis is not considered for statistical significance until an initial set of criteria is met. For example, it is common not to consider statistical significance of secondary outcomes unless the primary outcome reaches statistical significance. All simulations were performed in R, version 4.2.1.

## Noninferiority trials

Because noninferiority trials are used commonly to evaluate antibiotic use, and a cited advantage of DOOR/RADAR over noninferiority is sample-size efficiency, we performed power and sample-size calculations to compare estimates using DOOR/RADAR with those using the more established noninferiority trial design. We used failure rates and noninferiority margins that have been used in previous CAP studies.<sup>5</sup>

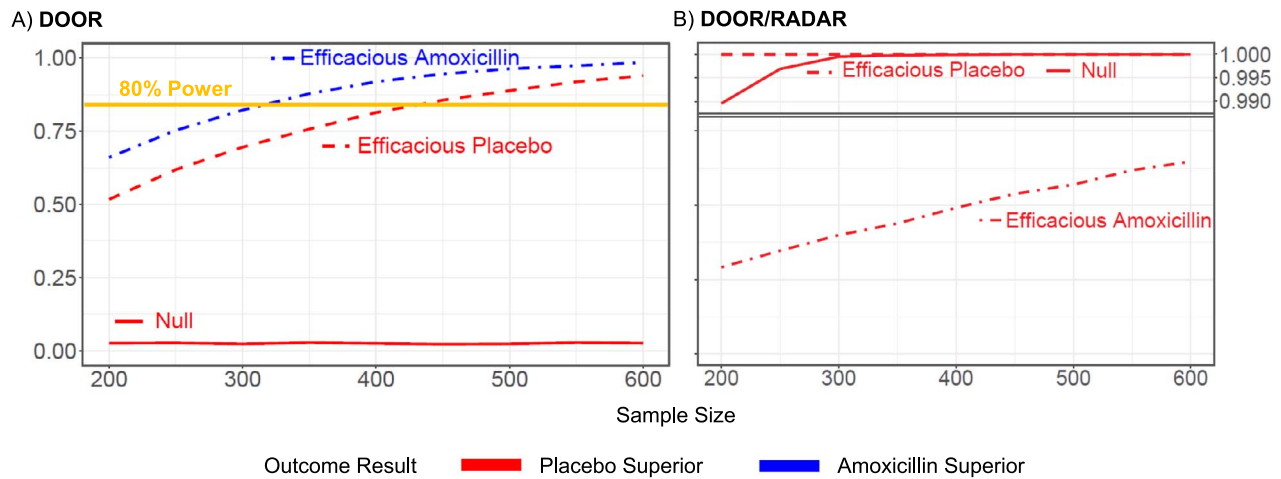
For these simulated trials, we assumed a dichotomous outcome of treatment failure (the most common outcome in prior CAP antibiotic trials) with the hypothesis that placebo was noninferior to amoxicillin with respect to treatment failure in young children with mild CAP who were treated as outpatients, a group in which viral etiologies predominate. We assumed equivalent failure rates between the placebo and amoxicillin arms and assessed failure rates between 2.5% and 10%.<sup>5</sup> Noninferiority margins ranged between an additional 2.5% and 5% above the failure rates, and a 1-sided type I error rate of 2.5% was used.

There are varied recommendations for noninferiority margins for clinical trials of CAP, mostly in adults. We based our noninferiority margins on prior literature and the recommendations from a Delphi consensus panel examining outcomes for clinical trials of children with mild CAP.<sup>21-23</sup> All noninferiority trial power calculations were performed with Power Analysis & Sample Size 2023 software (NCSS Statistical Software). Sample sizes were not inflated for anticipated loss to follow-up.

## Results

Power curves to detect differences between amoxicillin and placebo using DOOR or DOOR/RADAR for different trial conclusions are shown in Figure 1A and 1B, respectively. The null scenario, in which no differences exist in DOOR categories between arms (Figure 1A), shows the 1-sided type I error rate persevered at 2.5% when evaluating whether placebo was superior to amoxicillin. In the efficacious placebo scenario (scenario 2), the power to declare placebo was superior to amoxicillin and steadily increased from approximately 50% to 90% power when the total sample size was increased from 200 to 600. Approximately 400 patients were needed to achieve 80% power to detect a statistical difference between amoxicillin and placebo. In the efficacious amoxicillin scenario (scenario 3), amoxicillin was correctly identified as superior to placebo with 80% power with approximately 300 total patients using DOOR.

When considering DOOR/RADAR as the primary outcome, we had near 100% power to declare placebo was superior to amoxicillin for all sample sizes evaluated in simulation for the null and efficacious placebo scenarios (Figure 1B). Importantly, although using DOOR appropriately captured that amoxicillin was superior to placebo in the efficacious amoxicillin scenario,



**Figure 1.** Power curves for detecting differences among 3 simulated scenarios using A) desirability of outcome ranking (DOOR) and B) DOOR plus response adjusted for duration of antibiotic risk (RADAR) based on varying sample sizes. The color of the lines identifies the simulated trial conclusions or outcome results. For example, in the efficacious placebo scenario, a sample size of 250 will yield approximately 60% power to declare placebo superior to amoxicillin using DOOR and 99.7% power to declare placebo superior to amoxicillin using DOOR/RADAR.

when the DOOR/RADAR outcome was used, the trial conclusion was that placebo was the superior treatment. This occurred with less power compared with when DOOR matched the conclusions of DOOR/RADAR, ranging from approximately 30% to 65% with 200 and 600 trial participants, respectively.

We observed similar findings when evaluating probabilities of a better DOOR/RADAR outcome for a randomly selected individual (Table S2). For example, in scenario 1 using DOOR, we observed that patients receiving placebo had a 50% probability of a better DOOR outcome, which equated to the null situation being evaluated. In scenario 2, also as expected, patients receiving placebo had a greater than 50% probability of better DOOR (57%) and DOOR/RADAR (78%) outcomes. However, in scenario 3, once the DOOR/RADAR outcome was used, patients receiving placebo had more than a 50% probability of the better outcome (56%) even though DOOR appropriately concluded amoxicillin was efficacious (41%).

Similar findings were observed in the hypothetical example 2-arm trial of 2 different durations of antibiotic therapy (Figure S2). When we incorporated statistical gatekeeping into the analysis, we found the probability of a false conclusion in the efficacious amoxicillin scenario was greatly diminished (Table S3). In the specific scenarios evaluated, we found gatekeeping thresholds evaluating if the probability of a better DOOR being greater in the shorter duration arm was greater than 45% reduced this false conclusion to an acceptable level.

To allow for a feasibility comparison between an approach using the DOOR/RADAR outcome and an approach using noninferiority methodology, sample-size estimates for the noninferiority design are shown in Table 3. Sample-size estimates ranged from 300 participants, assuming a treatment failure rate of 2.5% and noninferiority margin of 7.5% with 80% power, up to 6056 participants, assuming a treatment failure rate of 10% and a noninferiority margin of 12.5% with 90% power.

## Discussion

Using statistical simulations, we quantified the power of DOOR/RADAR compared with noninferiority designs to detect significant differences in a hypothetical trial evaluating antibiotic vs no-antibiotic treatment strategies for mild CAP in children.

Although we found that the DOOR/RADAR design reduces the expected sample size at similar levels of power compared with a noninferiority design, we also found that DOOR/RADAR has a concerning potential to lead to an incorrect trial conclusion declaring placebo superior. Specifically, when considering a scenario in which a substantially higher rate of worsened clinical outcomes is expected in the placebo arm, the use of DOOR/RADAR can incorrectly identify placebo as the superior treatment.

The DOOR/RADAR outcome was created with a goal of combining efficacy, benefit, and safety risk into a single outcome.<sup>6</sup> Some investigators, however, have pointed out the inherent difficulties of interpreting results when these multidimensional concepts are collapsed for the sake of simplicity.<sup>24,25</sup> Investigators have raised several other important concerns regarding the DOOR/RADAR outcome. As our simulations demonstrate, the use of DOOR itself may lead to a large number of equivalent DOOR values between arms, even though 1 arm may have higher rates of worse clinical outcomes.<sup>12</sup> Other concerns include difficulty in the interpretation relative to the intention-to-treat principle<sup>13</sup> and treatment effect,<sup>13,15,16</sup> the possibility of obscuring important clinical differences between arms,<sup>12,13,15,17</sup> and the ability to manipulate the outcome by choosing different DOOR categories.<sup>13,16</sup> Concerns regarding the rank-based nature of DOOR being insufficient to detect trends in the most serious outcomes, especially in studies involving rare serious conditions, should be considered by clinicians and investigators designing DOOR ranks and in power calculations. Traditional noninferiority trials include these more serious outcomes as “failures,” reducing the likelihood of ignoring these infrequent, but extreme, cases. The participants who have adequate clinical responses are all grouped regardless of adverse event symptoms, which focuses the study on recovery. Similar to our findings, the use of DOOR/RADAR often permits substantially smaller sample sizes compared with traditional noninferiority study designs. The reduced sample size is an important advantage in clinical trials where there may be more challenges to recruit (eg, pediatrics and other vulnerable populations).

One of the fundamental issues related to DOOR/RADAR is the alignment of clinical/scientific hypotheses and statistical, or abstract, hypotheses. The traditionally accepted clinical trial paradigm involves defining a null hypothesis, collecting data from groups assigned different treatments, and assessing the



**Table 3.** Sample sizes required for various placebo and amoxicillin failure rates and non-inferiority thresholds to achieve 80% and 90% power.

| Placebo/amoxicillin failure rate <sup>a</sup> | Non-inferiority threshold | Sample size total (per arm) |             |
|---|---------------------------|-----------------------------|-------------|
|   |                           | 80% Power                   | 90% Power   |
| 2.5%  | 5.0%                      | 1222 (611)                  | 1644 (822)  |
| 2.5%  | 7.5%                      | 300 (150)                   | 428 (214)   |
| 5.0%  | 7.5%                      | 2380 (1190)                 | 3202 (1601) |
| 5.0%  | 10.0%                     | 594 (297)                   | 804 (402)   |
| 7.5%  | 10.0%                     | 3486 (1743)                 | 4666 (2333) |
| 7.5%  | 12.5%                     | 872 (436)                   | 1166 (583)  |
| 10.0%   | 12.5%                     | 4520 (2260)                 | 6056 (3028) |
| 10.0%   | 15.0%                     | 1126 (563)                  | 1514 (757)  |

<sup>a</sup>Failure rates are assumed to be the same between placebo and amoxicillin arms

likelihood of the observed outcomes under the null hypothesis. The challenges to this paradigm in trials that use DOOR/RADAR as an outcome become clear when one attempts to define a null hypothesis. Consider, as in our example, a 2-arm trial in which 1 arm receives antibiotics and the other receives placebo. At the population level, a true null hypothesis is one in which the expected outcome is the same in both arms. Because placebo receives a better ranking when clinical outcomes are tied, the null hypothesis of no difference in distribution between arms for DOOR/RADAR requires the amoxicillin arm to be more efficacious in DOOR categories in a way that the DOOR/RADAR outcome perfectly balances the overall distribution. Clinically, it is challenging to interpret such a null hypothesis so that when trial results reject the null hypothesis, it is unclear what is being rejected. At the individual patient level, the exercise becomes even more problematic because the null hypothesis is one in which a patient would have the same outcome, regardless of arm assignment. This is impossible, because having a different DOOR outcome is better or worse, and remaining within the same DOOR outcome category results in a better outcome with placebo. This reasoning can also be extended to trials comparing arms that represent different durations of antibiotic treatment.

Our study highlights the difficulty in identifying infrequent but harmful effects in a placebo arm for DOOR/RADAR that have not been previously discussed. These simulations demonstrate and quantify, under certain scenario parameters, the susceptibility of the DOOR/RADAR outcome to incorrectly recommend a clinically inferior treatment. In the hypothetical example of efficacious amoxicillin, the placebo group had an absolute increase in inadequate clinical response with additional emergency department or clinic visits of 7% (3% in the antibiotic group; 10% in the placebo group), an absolute increase in hospitalization of 8% (0.95% antibiotic group; 8.95% placebo group), and an absolute increase in rate of death of 4% (0.05% antibiotic group; 4.05% placebo group), yet DOOR/RADAR identified placebo as the superior treatment. This phenomenon is due to DOOR/RADAR favoring placebo for most patients who have an adequate clinical response (DOOR outcome categories 1-4) and having less weight, due to the total number in each category, for the small, but critical, distribution differences in the severe DOOR categories (DOOR outcome categories 6-8). Although the results of this hypothetical scenario are unlikely to occur in practice exactly as simulated, these results demonstrate an important potential drawback to using DOOR/RADAR.

The DOOR/RADAR outcome was recently used in a trial of 5 vs 10 days of antibiotic treatment for young children with mild CAP (SCOUT-CAP).<sup>18</sup> In that trial, most children fell into ranks 1

and 2 (of 8; 76% in the 5-days group; 78% in the 10-days group), signifying adequate clinical resolution with no or mild antibiotic-associated adverse effects. As observed in other CAP studies, fewer participants in the SCOUT-CAP trial experienced the most serious outcomes, which makes the trial results susceptible to the concealment of harmful effect findings discussed here. No participants died and few experienced severe events in SCOUT-CAP, indicating that the study's findings are likely valid; however, our results evaluating the DOOR/RADAR outcome overall suggest that caution is still required in making claims about infrequent, serious outcomes.

Operationally, when using DOOR/RADAR as the primary outcome, we recommend that researchers reduce this risk of concealing harmful treatment effects in shorter antibiotic-duration arms by establishing a statistically based gatekeeping approach. With DOOR/RADAR, study analyses may first consider the probability of a better DOOR outcome without the RADAR component<sup>6</sup> and continuing with DOOR/RADAR only if the observed probability of DOOR in the placebo group is at an acceptable level when compared with the antibiotic group (eg, > 50%, which reflects no worsening in DOOR). To create this threshold, investigators would need to determine acceptable differences in DOOR categories between arms and calculate what probability would be associated with this difference. Although we observed that a gatekeeping threshold of 0.45 reduced the probability of falsely claiming a shorter duration of treatment is superior using DOOR/RADAR, the acceptable threshold would likely depend on the exact population (eg, age) under study. Alternatively, DOOR/RADAR should be considered as a secondary or exploratory outcome to supplement the findings of the primary analysis.

Although using the DOOR/RADAR outcome can produce smaller required sample sizes compared with noninferiority trial designs, alternative outcomes might better support definite conclusions<sup>15</sup>; however, these alternative outcomes are typically compared using the noninferiority framework. The use of DOOR/RADAR could be beneficial in earlier phase studies to identify if a larger definitive trial is worthwhile.

In a simulated evaluation of the DOOR/RADAR outcome, we presented 3 scenarios to demonstrate power gains associated with increased sample sizes at numbers smaller than in comparative noninferiority trials. One scenario, however, highlights the limitations of further ranking patients by antibiotic-days (RADAR) that have the potential to reverse the findings of a study. We also did not present the clinically plausible scenario in which placebo performs slightly inferiorly to antibiotics in adequate clinical response, but in which placebo also results in fewer

side effects among participants with adequate clinical responses. Such a scenario would require determination of a placebo inferiority threshold acceptable to clinicians. Although there have been several versions of DOOR outcome categories proposed or evaluated for various conditions,<sup>16</sup> the simulations presented here only considered DOOR categories implemented in the SCOUT-CAP trial, given their relevance to pediatric CAP.<sup>18</sup> Additionally, the observed distribution of the SCOUT-CAP DOOR categories was used as the basis for the proposed antibiotic effect in our simulation scenarios, and additional baseline distributions were not considered. When simulating the number of antibiotic days, we used a binomial distribution, which produces a parametric relationship across days instead of allowing the distribution to have various peaks (eg, a group of patients who never take the antibiotic and a group of patients that take the full course).

## Conclusion

The use of both DOOR and DOOR/RADAR outcomes tends to reduce the required sample size compared with noninferiority studies in antibiotic studies of mild CAP in children. The DOOR/RADAR framework as an outcome in CAP trials that compares antibiotics vs placebo has the potential to reverse the clinical signal DOOR would provide. In scenarios that have the potential for this to occur, statistical gatekeeping assessing DOOR should be implemented prior to using the DOOR/RADAR framework.

## Supplementary material

Supplementary material is available at the *American Journal of Epidemiology* online.

## Funding

This work is supported by the National Heart, Lung and Blood Institute (grant R34-HL153474). PECARN is supported by the Health Resources and Services Administration of the US Department of Health and Human Services, in the Maternal and Child Health Bureau, under the Emergency Medical Services for Children (EMSC) program through the following cooperative agreements: EMSC Data Center, University of Utah (UJ5MC30824); Great Lakes Atlantic Children's Emergency Research (GLACiER), Nationwide Children's Hospital (U03MC28844); Hospitals of the Midwest Emergency Research Node (HOMERUN), Cincinnati Children's Hospital Medical Center (U03MC22684); Pediatric Emergency Medicine Northeast, West, and South (PEMNEWS), Columbia University Medical Center (U03MC00007); Pediatric Research in Injuries and Medical Emergencies (PRIME), University of California at Davis Medical Center (U03MC00001); Charlotte, Houston, and Milwaukee Prehospital (CHaMP) node, State University of New York at Buffalo (U03MC33154); Seattle-Texas-Los Angeles Research (STELAR), Seattle Children's Hospital (U03MC33156); and San Francisco-Oakland, Providence, Atlanta Research Collaborative (SPARC), Rhode Island Hospital/Hasbro Children's Hospital (U03MC49671).

## Conflict of interest

The authors declare no conflicts of interest.

## Disclaimer

This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by Health Resources and Services Administration, US Department of Health and Human Services, or the US government.

## Data availability

No patient-level data were used.

## References

1. Jain S, Self WH, Wunderink RG, et al. Community-acquired pneumonia requiring hospitalization. *N Engl J Med*. 2015; 373(24):2382. <https://doi.org/10.1056/NEJMc1511751>
2. Pfuntner A, Wier LM, Stocks C. Most frequent conditions in U.S. hospitals, 2011. In: *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*. Agency for Healthcare Research and Quality; 2013.
3. Bradley JS, Byington CL, Shah SS, et al. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the Pediatric Infectious Diseases Society and the Infectious Diseases Society of America. *Clin Infect Dis*. 2011;53(7):e25-e76. <https://doi.org/10.1093/cid/cir531>
4. Folgori L, Bielicki J, Ruiz B, et al. Harmonisation in study design and outcomes in paediatric antibiotic clinical trials: a systematic review. *Lancet Infect Dis*. 2016;16(9):e178-e189. [https://doi.org/10.1016/S1473-3099\(16\)00069-4](https://doi.org/10.1016/S1473-3099(16)00069-4)
5. Bradley JS, McCracken GH. Unique considerations in the evaluation of antibacterials in clinical trials for pediatric community-acquired pneumonia. *Clin Infect Dis*. 2008;47(Suppl\_3):S241-S248. <https://doi.org/10.1086/591410>
6. Evans SR, Rubin D, Follmann D, et al. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). *Clin Infect Dis*. 2015;61(5):800-806. <https://doi.org/10.1093/cid/civ495>
7. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med*. 2000;1(1):19-21. <https://doi.org/10.1186/cvm-1-1-019>
8. Powers JH, Cooper CK, Lin D, et al. Sample size and the ethics of non-inferiority trials. *Lancet*. 2005;366(9479):24-25. [https://doi.org/10.1016/S0140-6736\(05\)66817-1](https://doi.org/10.1016/S0140-6736(05)66817-1)
9. Powers JH. Noninferiority and equivalence trials: deciphering 'similarity' of medical interventions. *Stat Med*. 2008;27(3):343-352. <https://doi.org/10.1002/sim.3138>
10. Fleming TR. Current issues in non-inferiority trials. *Stat Med*. 2008;27(3):317-332. <https://doi.org/10.1002/sim.2855>
11. Evans S. Noninferiority clinical trials. *Chance*. 2009;22(53):56-62. <https://doi.org/10.1007/s00144-009-0032-z>
12. Molina J, Cisneros JM. Editorial commentary: a chance to change the paradigm of outcome assessment of antimicrobial stewardship programs. *Clin Infect Dis*. 2015;61(5):807-808. <https://doi.org/10.1093/cid/civ496>
13. Phillips PP, Morris TP, Walker AS. DOOR/RADAR: a gateway into the unknown? *Clin Infect Dis*. 2016;62(6):814-815. <https://doi.org/10.1093/cid/civ1002>
14. Giacobbe DR, Signori A. Interpreting desirability of outcome ranking (DOOR) analyses in observational studies in infectious diseases: caution still needed. *Eur J Clin Microbiol Infect Dis*. 2019;38(10):1985-1986. <https://doi.org/10.1007/s10096-019-03612-0>

15. Harris PNA, McNamara JF, Lye DC, et al. Proposed primary endpoints for use in clinical trials that compare treatment options for bloodstream infection in adults: a consensus definition. *Clin Microbiol Infect.* 2017;23(8):533-541. <https://doi.org/10.1016/j.cmi.2016.10.023>
16. Schweitzer VA, van Smeden M, Postma DF, et al. Response adjusted for days of antibiotic risk (RADAR): evaluation of a novel method to compare strategies to optimize antibiotic use. *Clin Microbiol Infect.* 2017;23(12):980-985. <https://doi.org/10.1016/j.cmi.2017.05.003>
17. Ong SWX, Petersiel N, Loewenthal MR, et al. Unlocking the DOOR-how to design, apply, analyse, and interpret desirability of outcome ranking endpoints in infectious diseases clinical trials. *Clin Microbiol Infect.* 2023;29(8):1024-1030. <https://doi.org/10.1016/j.cmi.2023.05.003>
18. Williams DJ, Creech CB, Walter EB, et al. Short- vs standard-course outpatient antibiotic therapy for community-acquired pneumonia in children: the SCOUT-CAP randomized clinical trial. *JAMA Pediatr.* 2022;176(3):253-261. <https://doi.org/10.1001/jamapediatrics.2021.5547>
19. Jain S, Williams DJ, Arnold SR, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. *N Engl J Med.* 2015;372(9):835-845. <https://doi.org/10.1056/NEJMoa1405870>
20. Pernica JM, Harman S, Kam AJ, et al. Short-course antimicrobial therapy for Pediatric community-acquired pneumonia: the SAFER randomized clinical trial. *JAMA Pediatr.* 2021;175(5):475-482. <https://doi.org/10.1001/jamapediatrics.2020.6735>
21. Bielicki JA, Stöhr W, Barratt S, et al. Effect of amoxicillin dose and treatment duration on the need for antibiotic re-treatment in children with community-acquired pneumonia: the CAP-IT randomized clinical trial. *JAMA.* 2021;326(17):1713-1724. <https://doi.org/10.1001/jama.2021.17843>
22. Florin TA, Melnikow J, Gosdin M, et al. Developing consensus on clinical outcomes for children with mild pneumonia: a Delphi study. *J Pediatric Infect Dis Soc.* 2023;12(2):83-88. <https://doi.org/10.1093/jpids/piac123>
23. Spellberg B, Talbot G. Recommended design features of future clinical trials of antibacterial agents for hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumonia. *Clin Infect Dis.* 2010;51(Suppl 1):S150-S170. <https://doi.org/10.1086/653065>
24. Ferreira-González I, Alonso-Coello P, Solà I, et al. Variables de resultado combinadas en los ensayos clínicos. *Rev Esp Cardiol.* 2008;61(3):283-290. <https://doi.org/10.1157/13116656>
25. Shaw PA. Use of composite outcomes to assess risk-benefit in clinical trials. *Clin Trials.* 2018;15(4):352-358. <https://doi.org/10.1177/1740774518784010>