In the study of the activity of PA-824 in mice performed by Nuernberger et al. (3), the authors observe that “regimens were not statistically different on the primary outcome” and propose that new studies using shorter regimens are needed “to determine conclusively whether substitution of PA-824 for isoniazid has the potential” to shorten treatment in humans. We agree but have concerns about the promotion of relapse to “primary endpoint” in the mouse model. We believe that most recent studies are underpowered for this kind of comparison.

Pharmacokinetic factors were carefully evaluated in the model reported, but there is currently less experience with relapse as a pharmacodynamic endpoint in this system compared for example with Cornell-type models where a fine balance of maturity of infection, duration of treatment, and subsequent immunosuppression markedly affects outcomes (5). Though the relapse rate can be conveniently manipulated by modifying experimental conditions, there are issues of both reproducibility and generalizability. Positive cultures on solid media may not reflect “persistor” organisms as well as liquid culture (1), while their timing and occurrence in otherwise healthy animals raise questions of how well this reflects the relapse process in humans.

For comparisons based on colony counting, the low fractional error of counts (~10%) means that sacrifice of as few as six mice per arm per time point is required to achieve 90% power for typical differences between regimens. By contrast, with 19 mice remaining per arm at 6 months, equal allocation, and a type I error of 5%, a power of even 80% for a superiority comparison requires an absolute difference of 40% in relapse rates between regimens, making this the minimum target rate in the control arm. In fact, this was the rate observed in this model for “standard” therapy at 4-months duration (4), similar to that in human trials of ultrashort-course regimens (2). If PA-824 reduced relapse at this duration to 0%, a modestly increased sample size would reasonably ensure that superiority was not missed. If it reduced relapse only to 20%, surely still an important result, more than 90 mice per arm would be required. Furthermore, this does not account for control of type I error, which dilutes the power of individual tests (to 60% for Bonferroni’s correction on six comparisons [6]), that we might require power greater than 80% in such pivotal animal studies (Table 1) or the cumulative sacrifices required for colony counts at intermediate time points. Clearly, superiority studies with 6-month therapy as comparator or at any reduced duration with control relapse rates of ≤10% may hardly be feasible.

The outstanding advantage of mouse models is that “sterilization” can in principle be directly quantified right up to the end of treatment. We question therefore whether less powerful categorical comparisons based on relapse are really more informative than those based on continuous measures. Even if we can accept the implicit analogy with human clinical trials, design factors should be carefully considered in undertaking mouse studies based on relapse if potentially important treatment effects are not to be dismissed.

### Table 1. Sample size per arm required to show superiority in relapse rate for a two-arm mouse study using equal allocation based on relapse (using Fisher’s exact test with \( \alpha = 0.05 \) and \( 1 - \beta = 0.90 \))

<table>
<thead>
<tr>
<th>Relapse rate in treatment arm</th>
<th>Relapse rate in the control arm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5</td>
</tr>
<tr>
<td>0</td>
<td>490</td>
</tr>
<tr>
<td>2.5</td>
<td>1,291</td>
</tr>
<tr>
<td>5</td>
<td>620</td>
</tr>
<tr>
<td>10</td>
<td>286</td>
</tr>
<tr>
<td>20</td>
<td>118</td>
</tr>
</tbody>
</table>

### References


**Authors’ Reply**

We thank Davies et al. for their interest in our recent study and appreciate the opportunity to respond to their thoughtful comments. We agree wholeheartedly that “design factors should be carefully considered in undertaking mouse studies based on relapse.” As we acknowledged, the design of this study prevented a determination of superiority for any of the...
experimental PA-824-containing regimens on the basis of relapse. However, the design of this particular study need not cast doubt on the use of relapse as an appropriate endpoint for the evaluation of experimental regimens in the murine model.

By analogy with the response to treatment in human tuberculosis, we assert that relapse remains the best measure of the sterilizing activity of an experimental regimen in the murine model. Experiments have repeatedly demonstrated that mice under treatment with effective combination regimens eventually reach a stage in which mice sacrificed for quantitative CFU counts in lung homogenates will have no detectable CFU on solid agar, but mice left untreated for 3 to 6 months may “relapse” spontaneously with detectable CFU. We see no reason to think that the persistently viable bacilli detected in such relapses are not representative of the population of persisters that lead to relapse of human disease. Therefore, we believe that measuring the activity of experimental regimens against this bacillary population through the assessment of relapse is essential to evaluate the potential of new regimens to shorten the duration of therapy. This is the same principle that underlies the use of the Cornell model that Davies et al. refer to.

Studies in the murine model should be designed with the power to detect meaningful treatment effects. When using relapse as an endpoint, this goal is best accomplished by restricting the number of experimental groups under study and selecting appropriate time points for assessment of relapse. Preliminary experiments based on serial CFU count determinations can eliminate combinations that are clearly less active than the control regimen and reveal the approximate time to culture negativity. Subsequent experiments can then be designed to compare selected regimens by assessing relapse at serial time points during treatment, including early time points at which at least 50% of control mice are expected to relapse and subsequent time points to determine the minimum duration of treatment with experimental regimens required for durable cure (i.e., no relapse at all). When the primary objective is to identify regimens that have the potential to significantly shorten the necessary duration of therapy (e.g., from 6 months to 4 months or less), we feel justified in requiring a minimum difference of 40 to 50% in the relapse rate compared to the standard 6-month regimen at a given time point. As Davies et al. demonstrate, studies with the power to detect such differences are quite feasible in the murine model.

This is an exciting time to be involved in drug development for tuberculosis, as there are at least five new drugs in clinical trials for this indication. This presents the fortunate dilemma of determining the optimal combination(s) of new and existing drugs to bring forward into phases II and III. Because the murine model is highly tractable and well validated through extensive experience with existing drugs, it will continue to play a pivotal role in evaluating novel drug combinations. Certainly, the potential of many such combinations can be assessed on the basis of quantitative CFU counts alone as promoted by Davies et al. However, it is ultimately necessary to demonstrate that the most promising novel regimens possess the expected sterilizing activity against persisters to support their use in shortened treatment regimens. The comments of Davies et al. serve as a timely reminder that experiments designed with relapse as an outcome require careful determination of sample size and timing of endpoints.

Eric L. Nuermberger*
Jacques H. Grosset
Center for Tuberculosis Research
Department of Medicine
Johns Hopkins University School of Medicine
1550 Orleans St.
Baltimore, Maryland 21231

*Phone: (410) 502-0580
Fax: (410) 614-8173
E-mail: enuermb@jhmi.edu