Heterogeneity of the Baseline Risk within Patient Populations of Clinical Trials

A Proposed Evaluation Algorithm

John P. A. Ioannidis¹ and Joseph Lau²

In this paper, the authors present an evaluation algorithm for systematic assessment of the observed heterogeneity in disease risk within trial populations. Predictive models are used to estimate the predicted patient hazards, the odds of having an event in the upper risk quartile (ODU) and the lower risk quartile (ODL), and the odds ratio (rate ratio for time-to-event analyses) for having an event in the upper risk quartile versus the lower risk quartile (extreme quartile odds ratio (EQuOR) and extreme quartile rate ratio (EQuRR)). The ranges for these metrics depend on the extent to which predictors of the outcome of interest exist and are known and the extent to which data are collected in the trial, as well as on the eligibility criteria and the specific patients who are actually enrolled. ODU, ODL, and EQuOR values are used to systematically interpret the results for patients at different levels of risk, to evaluate generalizability, and to determine the need for subgroup analyses. Individual data for five outcomes from three trials (n = 842, 913, and 1,001, respectively) are used as examples. Observed EQuOR values ranged from 1.5 (very little predicted heterogeneity) to 59 (large heterogeneity). EQuRR values ranged from 2 to 46. ODU values ranged from 0.24 to 3.19 (generally high risk), and ODL values ranged from 0.01 (clinically negligible risk) to 0.16 (clinically meaningful risk). The algorithm may also be used for comparing diverse trials (e.g., in meta-analyses) and used prospectively for designing future trials, as shown in simulations. Am J Epidemiol 1998;148:1117-26.

Individuals enrolled in a clinical trial have distinct characteristics leading to unavoidable heterogeneity in the patient population of any study. The extent and nature of this heterogeneity is determined by several factors. These include how restrictive (or unrestrictive) the eligibility criteria are, how diverse (or homogeneous) the population of potentially eligible patients is during the process of enrollment, which specific eligible patients actually become enrolled, and, most importantly, how diverse the studied disease process is itself. Diverse characteristics may result in diverse individual risks for the disease outcome being studied. It has been proposed that the efficacy of medical interventions shown by the results of each trial should be interpreted in the context of these individual patient risks (1-3). This is particularly important when the net benefit minus the potential harm from a treatment is not the same for patients at different levels of baseline risk (1). Evidence has been accumulating on several circumstances in which the treatment effect seems to depend on the baseline risk of the treated patients (4-11). In these cases, extrapolating the results of a clinical trial to all patients regardless of level of risk would be inappropriate.

Because of the importance of these issues, a standardized approach for evaluating and interpreting the baseline risk heterogeneity of individual patients within clinical trials is desirable. We propose an algorithm that can be used in any clinical trial to evaluate what is known of the diversity of the enrolled population in terms of the baseline risk of individual patients and what other considerations should be addressed given this measured heterogeneity in interpreting trial results.
ALGORITHM

The algorithm has the following four steps:

First, all predictors that are known to be potentially important for determining the risk of the studied outcome are considered. Data may not have been collected on all of them in a given trial. Past trials have collected variable amounts of information on the predictors of patient risk. Some trials have collected very detailed data on any potentially important variable, and have even stratified randomized patients according to some of these variables in order to avoid imbalance despite randomization. In some other cases, where simple data collection for large trials has been advocated (12) to reduce cost and increase convenience, information on several predictors may not be collected, even if these predictors are known to be potentially important determinants of disease risk. The predictors used in each study should be those for which there is already external evidence to suggest their importance from earlier trials, observational data, or known pathophysiology. Ideally, predictors should be prespecified before data analysis. If they are not prespecified (when there is little prior evidence from other observations regarding which biologic predictors may be important), the approach should be considered exploratory.

Second, a predictive model is built using the known available predictors for the patients enrolled in the study, and the predicted risk of each enrolled patient is estimated on the basis of the developed model. The preferred statistical model depends on the nature of the data. Therefore, it may be proportional hazards for time-to-event data, logistic regression for binary data, weighted linear regression for continuous outcomes, or discriminant analysis or polytomous logistic regression for outcomes with multiple possible states.

Third, a histogram showing the number of patients at each level of predicted risk is plotted in order to visualize the extent of known predicted heterogeneity in the trial population. The following metrics may also be estimated: 1) the odds of having an event of interest for patients in the upper quartile of predicted risk (ODU) and, for time-to-event analyses, the odds of having an event of interest in the early quarter of follow-up for patients in the upper quartile of predicted risk (ODUE) (the selection of the first quarter is not absolute but is simply selected for standardizing all of the analyses); 2) the odds of having an event of interest for patients in the lower quartile of predicted risk (ODL); and 3) the extreme quartile odds ratio (EQuOR), defined as the odds of having an event of interest if a patient belongs in the upper quartile of predicted risk versus the lower quartile. For time-to-event analyses, the respective metric is the extreme quartile rate ratio (EQuRR).

Fourth, the results are interpreted on the basis of these metrics. The EQuOR (or EQuRR) gives an estimate of how diverse the predicted baseline risk is among patients in the two outer quartiles and is thus a measure of the known measured diversity. The ODU and ODL give estimates of whether this diversity is contributed mostly by high-risk patients, mostly by low-risk patients, or about equally by both groups. Depending on the value of these metrics, there are different possibilities to be considered for further analyses and for interpretation of the trial results.

Below, the EQuORs are classified into three categories for convenience of discussion, but it should be remembered that these metrics are continuous variables and the cutoff values are not absolute. The categorization is only used here to facilitate the presentation of some key points regarding the interpretation of the metric. The suggested cutoff values are set to contain fivefold differences within each category. A similar classification may be used for EQuRRs. EQuRR values may tend to be smaller than EQuORs, since the rate ratio is not identical to the odds ratio.

Low range EQuOR. The interpretation of a low range EQuOR (typically 1–5) is that there is little heterogeneity based on known risk predictors in the trial population. Possibilities include 1) data on known predictors of disease risk were not collected by the trial; 2) the trial used narrow eligibility criteria or enrolled a very homogeneous patient population, or the disease process is very homogeneous; 3) predictors of disease risk are still unknown or the predictive model is not adequate; 4) a combination of the above; or 5) the disease process is largely unpredictable (no strong risk factors exist). The implications are that the generalizability for patients at different levels of risk cannot be assessed. Even if predictors are known and the enrolled trial population was purposefully made homogeneous, strictly the results apply only to such patients as those enrolled; extrapolation to other populations may not be warranted (13). If the disease process is not very homogeneous, the patient population was not made purposefully homogeneous, and predictors are unknown, more research is needed for understanding better the disease process. The trial results could easily have been affected by chance if important unknown or uncollected predictors were allocated unevenly between arms by chance (13). Large sample sizes are needed to obviate this possibility through randomization.

Moderate range EQuOR. The interpretation of a moderate range EQuOR (typically 5–25) is that there is substantial heterogeneity based on known risk pre-
dictors in the trial population. There are two major possibilities to consider alone or in combination: First, influential predictors were known and data on them were collected by the trial, but it is very likely that there are additional unknown or uncollected predictors, unless the trial eligibility criteria were narrow or a very homogeneous population was enrolled in the trial. Second, unrestrictive eligibility criteria were used and either 1) some patients without any meaningful risk or even without the disease were enrolled or 2) some patients with a very high risk, possibly even representing a different disease process, were enrolled. The implications are that it may be important to address treatment-risk interactions in order to evaluate the consistency of the treatment effect for patients at different levels of risk and subgroup analyses (2, 3) and the clinical meaning of the treatment effect may be evaluated for patients at different levels of risk (1). If trial eligibility criteria were not meant to be narrow, further research is needed to identify potential additional risk factors. Finally, it is worthwhile to evaluate by inspecting the ODL and ODU (or ODUE) whether, either because aspects of the disease process were unknown or because eligibility criteria were poorly defined, either a large proportion of patients without any risk or a large proportion of patients with very high risk were enrolled in the trial.

High range EQuOR. The interpretation of a high range EQuOR (typically >25) is that there is large heterogeneity based on known risk predictors in the trial population. There are two major possibilities to consider, alone or in combination: First, very influential predictors are known and data on them have been collected in the trial, even if unknown predictors may still exist. Second, very unrestrictive eligibility criteria were used; many patients without any meaningful risk or even without the disease or many patients with a very high risk (possibly even representing a different disease process) were enrolled. The implications are that it is important to address treatment-risk interactions to evaluate the consistency of the treatment effect for patients at different levels of risk and subgroup analyses (2, 3) and the clinical meaning of the treatment effect must be evaluated for patients at different levels of risk. Even if the treatment effect, expressed as the relative risk, is the same at all levels of risk, the absolute effect for some patients may be clinically meaningless (1). It is necessary to evaluate by inspecting the ODL and ODU (or ODUE) whether, either because aspects of the disease process were unknown or because eligibility criteria were very unrestrictive, either a large proportion of patients without any risk at all or a large proportion of patients with extremely high risk were enrolled in the trial. Finally, the search for additional, yet-unknown predictors should not be abandoned or discouraged because of very high EQuOR values.

**CLINICAL TRIAL EXAMPLES**

**Trial databases and model building**

We performed analyses of five endpoints from three randomized trials of human immunodeficiency virus infection to illustrate the principles of the algorithm. These trials were AIDS Clinical Trials Group (ACTG) 081 (14), ACTG 116B (15), and ACTG 155 (16). ACTG 081 (14) was a trial of 842 patients comparing three strategies starting with trimethoprim-sulfamethoxazole, dapsone, or aerosolized pentamidine for the primary prophylaxis of *Pneumocystis carinii* pneumonia—the primary endpoint being prophylaxis failure. ACTG 116B (15) was a trial of 913 patients comparing the continuation of zidovudine with a switch to didanosine at one of two different doses in human immunodeficiency virus-positive patients with either symptomatic disease and <300 CD4 cells per mm$^3$ or asymptomatic disease and <200 CD4 cells per mm$^3$—the primary endpoint being progression to acquired immunodeficiency syndrome or death and the secondary endpoint being survival. ACTG 155 (16) was a trial of 1,001 patients comparing the continuation of zidovudine or a switch to zalcitabine with a zalcitabine/zidovudine combination in human immunodeficiency virus-infected patients with either symptomatic disease and <300 CD4 cells per mm$^3$ or asymptomatic disease and <200 CD4 cells per mm$^3$—the primary endpoint being progression to acquired immunodeficiency syndrome or death and the secondary endpoint being survival.

In all cases, we considered all of the a priori documented predictors that had been collected and recorded for at least 90 percent of patients in each trial in a multivariate Cox proportional hazards model for each endpoint of interest (17). These variables typically included patient age, performance status (Karnofsky score), clinical disease status (presence or absence of acquired immunodeficiency syndrome, symptomatic or asymptomatic disease), CD4 cell count (the square root was used in the model to improve model fit), the presence or absence of detectable p24 antigen (>25 pg/ml), the use of prophylaxis for *P. carinii*, hemoglobin level, and body weight. Not all variables had had data collected or were pertinent in all data sets. The multivariate model was built with forward stepwise selection of the available variables using a log likelihood ratio criterion for the selection process, and with variables entered at $p < 0.05$ and
removed at \( p > 0.10 \) (17). Interaction terms between the selected predictors were evaluated in the final model, but none were important enough to consider, and, except for the square root transformation of CD4 cell count, no other nonlinear transformations seemed important upon inspection of univariate analyses. Coefficients of determination were also calculated (18, 19).

**Heterogeneity metrics**

Table 1 shows the predictive models for each considered endpoint in each trial, and table 2 shows the estimated values of odds in different risk quartiles, as well as the heterogeneity metrics. Values of EQuOR range from 1.5 (very little predicted heterogeneity) to 59 (large predicted heterogeneity), and the respective range for EQuRR is 2 to 46. ODU values range from 0.24 to 3.19 (generally high risk), and ODUE values range from 0.06 to 0.58 (generally substantial risk, but far from the claim that these patients universally suffered an event in the early period of follow-up). ODL values range from 0.01 (clinically negligible risk) to 0.16 (clinically meaningful risk). Figure 1 shows histograms for the hazard of the event of interest for each of the analyses. There is minimal predicted heterogeneity for the risk of *P. carinii* pneumonia in ACTG 081, moderate heterogeneity for disease progression in ACTG 116B and ACTG 155, and larger heterogeneity for the mortality risk in ACTG 116B and ACTG 155.

**Interpretation and further considerations**

ACTG 081 has a very low EQuOR and EQuRR. The only predictor of *P. carinii* pneumonia in the trial was CD4 cell count (14). This was the only known influential predictor of *P. carinii* pneumonia risk at the time the study was launched. The study narrowed eligibility within a range of CD4 cell count where risk would be substantial. Subsequent research has identified some other predictors of risk of *P. carinii* pneumonia, including viral load (20) and possibly smoking (21), but the specific determinants of *P. carinii* pneumonia risk are still unknown to a large extent. The implications of a low EQuOR are that the generalizability of the trial results, which showed no significant

<table>
<thead>
<tr>
<th>Trial</th>
<th>Endpoint</th>
<th>Predictors in multivariate model</th>
<th>Coefficient</th>
<th>Predictor</th>
<th>Hazard ratio</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTG 081</td>
<td><em>Pneumocystis carinii</em> pneumonia</td>
<td>CD4 cell count (per μl; square root)</td>
<td>-0.0814</td>
<td></td>
<td>0.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ACTG 116B</td>
<td>AIDS/death</td>
<td>Hemoglobin (mg/dl)</td>
<td>-0.1572</td>
<td></td>
<td>0.85</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS diagnosis</td>
<td>0.3717</td>
<td></td>
<td>1.45</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Karnofsky score</td>
<td>-0.0290</td>
<td></td>
<td>0.97</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD4 cell count (per μl; square root)</td>
<td>-0.1336</td>
<td></td>
<td>0.87</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ACTG 116B</td>
<td>Death</td>
<td>Age (years)</td>
<td>0.0318</td>
<td></td>
<td>1.03</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin (mg/dl)</td>
<td>-0.2554</td>
<td></td>
<td>0.78</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIDS diagnosis</td>
<td>0.8639</td>
<td></td>
<td>2.37</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Karnofsky score</td>
<td>-0.0287</td>
<td></td>
<td>0.97</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD4 cell count (per μl; square root)</td>
<td>-0.1666</td>
<td></td>
<td>0.85</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ACTG 155</td>
<td>AIDS/death</td>
<td>Age (years)</td>
<td>0.0219</td>
<td></td>
<td>1.02</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p24 antigenemia (&gt;25 pg/ml)</td>
<td>0.2705</td>
<td></td>
<td>1.31</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Symptoms</td>
<td>0.3636</td>
<td></td>
<td>1.44</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No prior need for <em>P. carinii</em> prophylaxis</td>
<td>0.7166</td>
<td></td>
<td>0.49</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD4 cell count (per μl; square root)</td>
<td>-0.1673</td>
<td></td>
<td>0.85</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weight (kg)</td>
<td>-0.0117</td>
<td></td>
<td>0.99</td>
<td>0.001</td>
</tr>
<tr>
<td>ACTG 155</td>
<td>Death</td>
<td>Age (years)</td>
<td>0.0393</td>
<td></td>
<td>1.04</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p24 antigenemia (&gt;25 pg/ml)</td>
<td>0.3924</td>
<td></td>
<td>1.48</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Symptoms</td>
<td>0.7162</td>
<td></td>
<td>2.04</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD4 cell count (per μl; square root)</td>
<td>-0.2217</td>
<td></td>
<td>0.80</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weight (kg)</td>
<td>-0.0166</td>
<td></td>
<td>0.98</td>
<td>0.027</td>
</tr>
</tbody>
</table>

* For details on the development of the multivariate Cox model, see text. Data on all of the considered predictors were available for 835 of 842 patients enrolled in ACTG 081 (14), 906 of 913 patients randomized in ACTG 116B (15), and 925 and 917 of 1,001 patients randomized in ACTG 155 (16) for the AIDS/death and death endpoints, respectively.

† ACTG, AIDS Clinical Trials Group; AIDS, acquired immunodeficiency syndrome.
difference between the three compared regimens, cannot be assessed. It is unknown whether the results would be applicable to different populations. A meta-analysis (9) and a decision analysis (22) suggest that the choice of optimal regimen may be influenced by the risk of the disease. Further research is recommended in order to define better determinants of P. carinii pneumonia risk, even if some of the disease process may be unpredictable and chaotic to a certain extent—e.g., exposure of an individual to the pathogen may largely be a random process.

Several important predictors of risk were available for both ACTG 116B and ACTG 155. All influential predictors known at the time at which the studies were launched were considered for data collection by these trials. Both studies have moderately high EQuOR and EQuRR values with regard to progression to acquired immunodeficiency syndrome or death and even higher values with regard to mortality. This means that their patient populations are largely heterogeneous according to known predictors. The larger heterogeneity seen when death is considered may be due to the fact that there may exist larger true variability for time to death (a late, final event) than for disease progression (an earlier, more proximal event). The ODL values for disease progression in both trials are not negligible, meaning that even low-risk patients have a meaningful risk of disease: The trial population was not contaminated by patients at no risk of disease, and the results are clinically meaningful even for low-risk patients. However, if claims are to be made for survival only, then both trials (particularly ACTG 155) have very low ODL values for this endpoint. Strictly speaking, the trial results for survival are clinically uninterpretable for a large portion of the studied patients, regardless of whether or not an overall relative benefit is detected. The ODU is high for both disease progression and survival, but the ODU is not too high to suggest that the trial included patients who would have promptly progressed and died regardless of any treatment. The high EQuOR in both trials suggests that it is important to assess treatment-risk interactions and perform subgroup analyses according to known risk factors in order to evaluate the potential for heterogeneity of the treatment effect and its clinical meaning in different subgroups.

In fact, both ACTG 116B and ACTG 155 reported that a treatment benefit for disease progression was seen only for specific subgroups of patients. In ACTG 116B, only patients without acquired immunodeficiency syndrome were reported to benefit from didanosine (15). ACTG 155 reported a significant benefit for the combination regimen only in patients with CD4 cell counts above 150 per mm³, and there was a clear trend for decreasing benefit with lower CD4 cell counts (16). In both trials, in models considering treatment assignment and baseline risk, the interaction terms were of substantial magnitude, even if not formally statistically significant due to the limited sample size. For the primary endpoint in ACTG 116B, the didanosine versus zidovudine comparison hazard ratios in the four risk quartiles (lower to higher) were 0.48 (95 percent confidence interval (CI) 0.18–1.13), 0.68 (95 percent CI 0.35–1.30), 0.79 (95 percent CI 0.47–1.35) and 0.84 (95 percent CI 0.46–1.54). In ACTG 155, the respective quartile hazard ratios for the combination therapy versus monotherapy comparison were 0.66 (95 percent CI 0.28–1.52), 0.66 (95 percent CI 0.37–1.17), 0.79 (95 percent CI 0.47–1.34),

### TABLE 2. Within-trial heterogeneity metrics from proportional hazards predictive models and model fit for five endpoints in three different clinical trials*

<table>
<thead>
<tr>
<th>Protocol (endpoint)</th>
<th>Observed (predicted) event odds in predicted risk quartile</th>
<th>Observed ODUE†</th>
<th>EQuOR†</th>
<th>95% CI§</th>
<th>EQuRR†</th>
<th>95% CI</th>
<th>$R^2$ (1 - R²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTG 081 (PCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTG 116B (AIDS/§/death)</td>
<td>0.24 (0.24) 0.21 (0.20) 0.17 (0.19) 0.16 (0.14)</td>
<td>0.07</td>
<td>1.5</td>
<td>0.9–2.5</td>
<td>2.0</td>
<td>1.2–3.3</td>
<td>0.02 (0.02)</td>
</tr>
<tr>
<td>ACTG 116B (death)</td>
<td>2.20 (2.39) 0.80 (0.66) 0.30 (0.29) 0.10 (0.16)</td>
<td>0.55</td>
<td>23</td>
<td>13–39</td>
<td>13</td>
<td>8–21</td>
<td>0.22 (0.22)</td>
</tr>
<tr>
<td>ACTG 155 (AIDS/death)</td>
<td>0.73 (0.79) 0.25 (0.21) 0.06 (0.07) 0.02 (0.03)</td>
<td>0.11</td>
<td>41</td>
<td>15–114</td>
<td>30</td>
<td>11–80</td>
<td>0.19 (0.22)</td>
</tr>
<tr>
<td>ACTG 155 (death)</td>
<td>3.19 (2.90) 0.76 (0.90) 0.47 (0.40) 0.14 (0.16)</td>
<td>0.58</td>
<td>23</td>
<td>14–37</td>
<td>12</td>
<td>8–18</td>
<td>0.26 (0.26)</td>
</tr>
<tr>
<td>ACTG 116B, ACTG 155</td>
<td>0.80 (0.76) 0.16 (0.19) 0.11 (0.08) 0.01 (0.03)</td>
<td>0.06</td>
<td>59</td>
<td>18–190</td>
<td>46</td>
<td>15–146</td>
<td>0.19 (0.20)</td>
</tr>
</tbody>
</table>

* AIDS Clinical Trials Group (ACTG) 081 (14), ACTG 116B (15), and ACTG 155 (16).
† ODU, odds of event of interest in the upper risk quartile during the first quarter of follow-up; EQuOR, extreme quartile odds ratio, calculated as the ratio of the odds of an event of interest in the upper risk quartile (ODU) to the odds of an event of interest in the lower risk quartile (ODL); EQuRR, extreme quartile rate ratio, calculated from a proportional hazards model in which only patients in the extreme quartiles are considered and quartile membership is used as a dummy variable (0 = lower risk quartile, 1 = upper risk quartile).
‡ $R^2$, coefficient of determination, calculated as $1 - \exp[\text{change in } -2\text{LL}/n]$, where LL is the log likelihood and n is the number of patients in the analysis.
§ CI, confidence interval; ACTG, AIDS Clinical Trials Group; PCP, Pneumocystis carinii pneumonia; AIDS, acquired immunodeficiency syndrome.
and 1.04 (95 percent CI 0.56–1.92). A subsequent meta-analysis of several trials of antiretroviral changes in treatment-experienced patients confirmed the trend of decreasing benefit in more advanced disease stages (8). Finally, even such high EQuOR values do not mean that all important predictors were known when the trials were launched. For example, subsequent research proved the importance of human immunodeficiency virus RNA load as an independent predictor of disease progression and survival (23–25).

Treatment assignment was generally not included in these models. In most cases, the relative treatment effect in large trials is likely to be of relatively small magnitude, and predictors of risk should not be different in treated patients than in untreated patients. Sensitivity analyses, including or not including treatment assignment, may need to be performed if the treatment effect is of sizable magnitude and/or predictors of risk behave very differently in treated and untreated patients. Cumulative treatment effects resulting from the additive effects of several medical interventions for the treatment of a disease may be of a larger magnitude, and may need to be considered when extrapolating the predicted risks from an early study to a new study in a different therapeutic background where the management of a disease may have improved substantially.
Distinguishing clinical heterogeneity metrics from goodness of fit and explained variance

Table 2 also shows the values of the coefficient of determination, \( R^2 \), for the presented predictive models. \( R^2 \) measures the percentage of the variability explained by each model for the respective data. Values range from 0.02 to 0.26 in the presented examples, as it is typical for predictive models to explain only a modest portion of the observed variability. \( R^2 \) is strictly a statistical measure of explained variability, and it offers no information on how extensive the predicted heterogeneity is or on its clinical importance. The same holds true for other "goodness-of-fit" metrics (or more accurately described, explained variance metrics), such as the area under the receiver operating characteristic curve, which is easier to interpret for binary outcomes in logistic regression but is less accurate for time-to-event analyses with substantial censoring (as in human immunodeficiency virus trials). In the same trial, adding important predictors to the model may tend to increase both EQuOR (or EQuRR) and explained variance metrics. However, the latter are not clinically interpretable and cannot be used to compare different trial populations, even with the same predictors. As an extreme example, \( R^2 = 1 \) can be obtained both for a population with EQuOR = 1 (a most homogeneous population in which the times-to-event are identical—thus perfectly predictable) and for a population with EQuOR = 100 (an extremely heterogeneous population of size \( N \) accurately predicted with \( \leq N \) variables). The coefficients of determination in ACTG 155 and ACTG 116B are fairly similar for the disease progression and mortality outcomes; if anything, for ACTG 155 a slightly larger percentage of the variability might be explained for disease progression than for mortality, even with the same predictors considered (age, symptoms, p24 antigen, weight, and square root of CD4 cell count), while EQuOR would still be smaller for disease progression than for mortality. This is because the variability for disease progression is less than the variability for mortality, probably because death is a later event that could be affected by more factors than the more proximal disease progression. Finally, "goodness-of-fit" metrics offer no clinical information on the magnitude of the risk for high- and low-risk patients, which ODU and ODL do.

Application to the design of future trials

In designing a trial, not only the expected event rates but also the expected heterogeneity range needs to be considered. The former affects sample size determinations, but the latter may be more important for generalizability. Information on the distribution of predicted risks from previously conducted trials may be used to assess whether the expected heterogeneity of the designed trial population seems appropriate to the investigators for the purposes of the new study. New information may also be incorporated in the simulations. Table 3 shows the predicted EQuRR values for simulated trials after modification of the eligibility criteria of ACTG 116B by restriction of either the CD4 range and/or acquired immunodeficiency syndrome status. Also shown are examples in which mixtures of populations are contemplated—such as the situation where a trial is being considered for implementation in different countries (e.g., the United States and developing countries) where event rates for patients with identical biologic predictors may vary substantially. Such differences may be due to differential availability of new, highly active therapies, varying adequacy of medical care, differential exposure to pathogens, or different epidemic phases (4). In these simulations, the predicted EQuRR values range from 3.7 to 45. Accordingly, the investigators may wish to put different emphasis on subgroup analyses and treatment-risk in-
teractions with different designs and endpoints. Subgroup analyses should have a more prominent place in designs of large heterogeneity. Anticipation of large heterogeneity in the simulations of the study design should lead to advance specification of subgroup analyses pertaining to the baseline risk. Simulations may also be contrasted against the final results of the new trial, since the population actually enrolled may differ from that expected.

**DISCUSSION**

We have presented an evaluation algorithm with which to assess and interpret the known predicted heterogeneity of disease risk among individual participants in clinical trials, with examples drawn from human immunodeficiency virus-related studies. The approach can be applied across clinical trials regardless of discipline. The algorithm may be particularly helpful in the clinical interpretation of the study design and results, in the comparison of different trials on the same subject, and in the design of future studies. It is easy to implement and may provide a standardized common framework with which to interpret the known predicted heterogeneity within randomized studies.

Clinical trials have typically tried to provide a definitive answer to the question of whether or not a new medical intervention "works." Although this approach will often lead to acceptable and generalizable answers, in many cases a single treatment effect may not apply to all individuals (1–3). Clinical trial experience has accumulated several such examples (4–11). Sometimes patients at different levels of risk experience different levels of benefit or harm from a treatment. In other cases, variables associated with differences in the treatment effect may not necessarily be correlated with differences in the disease risk as well. However, even in such cases, it is still essential to be able to interpret the treatment benefit or harm in relation to the predictable baseline risk a patient is facing. Depending on the nature of the benefit or harm, different physicians and patients may make different therapeutic decisions for different thresholds of known risk. These thresholds may not be the same for everyone: Some patients and physicians may only accept very conservative risk-benefit ratios, while others may be willing to risk more. In either case, it would be essential to know how heterogeneous was the population which was enrolled in a trial and how wide was the range of disease risks involved on the basis of known and recorded predictors of the disease before the results can be extrapolated and applied to clinical practice. Of course, generalizability may not depend only on factors reflected in the risk of the studied outcome, but the risk is likely to be a very important factor in medical decisions made at the individual level.

A study with a narrow range of individual risks cannot claim generalizability. In the absence of other evidence, the results of such trials are often extended to many other patient populations and circumstances, but the possibility of incorrect extrapolation is always present. In some trials, even if the entry criteria were unrestrictive, the predicted range of risks may be narrow only because the trial did not collect data on important predictors. Simple trials with large numbers of participants may claim generalizability (12), but one cannot exclude the possibility that diversity in the treatment effect across subpopulations has been missed (13). Simple large trials (12) may become less attractive as better understanding of the biologic components of disease processes in the biomedical sciences reveals more potential sources of patient heterogeneity.

On the other hand, a study with a very wide range of individual risks may be more generalizable, especially if large numbers of patients have been randomized, but its clinical interpretation requires caution. Relative treatment benefits and harms, expressed as relative risks, may vary in different subgroups of patients. Even if they are the same, their absolute magnitude (the risk difference) may make the treatment highly desirable for some high-risk patients but meaningless for low-risk patients. Subgroup analyses may have a higher yield in trials with large heterogeneity of patient risks.

By modeling heterogeneity in advance, our algorithm allows researchers to determine a priori which trials should be particularly targeted for subgroup analyses. Establishing which predictors should be considered in the model may often be a separate research undertaking in its own right. We suggest that the set of potential predictors should be prespecified before the trial analysis on the basis of prior evidence. When no such evidence exists or the existing evidence is considered unreliable, the algorithm may still be applied as an exploratory approach, but it should be specified as such; the resulting predictive models would then require validation in other data sets. Post hoc analyses may be more credible if they are consistent and have biologic plausibility.

Aside from assessment of the study design and clinical interpretation of the results, the proposed algorithm can also be used to compare the populations of different trials studying the same or similar questions on the basis of the same set of known predictors. This may be particularly helpful in the synthesis of such data through meta-analysis (26), as well as in the evaluation of the resulting pooled populations in meta-
Risk Heterogeneity within Clinical Trials

analyses of individual patient data. For such comparisons, the same predictors should be used in all of the compared trials. Even in situations where eligibility criteria and predictors seem straightforward, the algorithm may show more clearly what the composition of the enrolled population is, as compared with direct inference from the eligibility criteria. Trials with seemingly similar eligibility criteria may enroll different populations if they enroll mostly patients with different values for various predictors within the range of eligible values.

The proposed algorithm is different from the traditional approach of specifying particular subgroup contrasts on the basis of selected variables within trials. Subgroup analyses are often viewed with justifiable skepticism given their low power and the difficulty of validating them (27, 28). Often biologic plausibility is evoked post hoc to fit the data. The proposed algorithm, by contrast, uses prespecified variables that are already known to affect disease outcome in order to generate predictive scores and assess their range within the trial population. Such a standardized approach may offer uniformity in a field which has suffered from lack of a systematic approach. Even when only one or two important risk factors are known, the distribution of values for these predictors may be different in different trials, and the proposed algorithm offers a perspective of this scatter. For most diseases, past knowledge of risk factors had been limited (29), but the situation is changing rapidly, given the current progress in the molecular sciences. Human immunodeficiency virus infection offers such an example, as predictive power has improved dramatically over the past 15 years. The proposed approach offers advantages over subgroup analyses in such multifactorial prediction situations. Composite estimates of risk are more appropriate for decision-making at the individual patient level as compared with isolated subgroup analyses which ignore other predictors. The tradeoff is that in some situations the magnitude of the treatment effect may be specifically related to a single biologic predictor; the inclusion of this predictor in a multivariate prediction score model may blunt this association. Such associations still need to be addressed with traditional subgroup analyses, but ideally these should be prespecified, with a strong biologic rationale to support them.

Furthermore, these considerations may be extended to the design of future trials as well. Simulations may make a direct impression on the clinicians designing the study if the expected amount of diversity makes clinical sense or if they are unknowingly designing a study that is too restrictive or too heterogeneous. This approach would help investigators to decide a priori whether or not subgroup analyses should have a primary role in the trial (if a very heterogeneous trial is being contemplated). Finally, the expected heterogeneity may then also be compared with the observed heterogeneity after a study is conducted, to examine how much the enrolled population differed from that which the clinicians originally expected.

Some technical caveats should be noted. First, the proposed heterogeneity metrics presented here are largely arbitrary in that quartiles were selected. Similar metrics could be estimated for quintiles, deciles, and so forth. Continuous representation of the population risk distribution is most objective, and it could be combined with a description of actual event rates in small subsegments of the population. Nevertheless, the use of quartiles is practical, since it avoids comparison of subsegments containing very few patients (and even fewer events) and still refers to the extremes of the population.

Second, the algorithm should lead to several different considerations at each level of heterogeneity. This means that there is no single interpretation every time; rather, different possibilities need to be considered systematically.

Third, the observed heterogeneity will depend also on the precision with which predictors and events of interest are measured. Odds ratios may be increasingly attenuated (30, 31) and the predicted heterogeneity may be blunted in the presence of increasing measurement error. However, this is unlikely to be a problem affecting the clinical interpretation of the results, since usually the same measurement errors that operate in a clinical trial are likely to operate in clinical practice to the same degree or even a larger degree. For example, the measurement error in CD4 cell counts in a trial-registered laboratory is likely to be the same or slightly less than the measurement error in a community laboratory.

Finally, the clinical metrics should not be confused with explained variance and model fit. In that regard, appropriate model-building is also essential for obtaining useful results (19, 32). Predictive rules need to have both good discrimination and good calibration to be useful. Both are likely to improve as more predictors of disease are measured more accurately.

The proposed algorithm is easy to implement, and it may be used routinely in clinical trials of diverse disciplines. Accumulating experience from diverse trial databases may offer information on the evolution of predicted risk heterogeneity across trials within a medical discipline and on the performance and validation of trials with different levels of heterogeneity. Current advances in biology allow the design of low-heterogeneity trials in highly specified patient popula-
tions with strictly defined biologic characteristics. Initial proof-of-concept trials may benefit from being performed in biologically targeted high-risk patients with high absolute benefits. At the other end of the spectrum, the advent of meta-analyses has generated pooled analyses that probably represent mostly high-heterogeneity designs, since diverse studies are added together (26). These two major advances in clinical research seem to be pulling us in opposite directions. Meta-analytic approaches based on meta-regression analyses have been developed to understand study-level risk versus benefit (11, 33). Such approaches are helpful, but they can be affected by ecological fallacies (34, 35). The proposed algorithm focuses on individual-level risk and may be more suitable for studying systematically the merits and disadvantages of high and low predicted heterogeneity clinical research at the individual patient level.

REFERENCES