Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials

James F. Burke, MD, MS; Rodney A. Hayward, MD; Jason P. Nelson, MPH; David M. Kent, MD, MS

Background—Recent proposals suggest that risk-stratified analyses of clinical trials be routinely performed to better enable tailoring of treatment decisions to individuals. Trial data can be stratified using externally developed risk models (eg, Framingham risk score), but such models are not always available. We sought to determine whether internally developed risk models, developed directly on trial data, introduce bias compared with external models.

Methods and Results—We simulated a large patient population with known risk factors and outcomes. Clinical trials were then simulated by repeatedly drawing from the patient population assuming a specified relative treatment effect in the experimental arm, which either did or did not vary according to a subject’s baseline risk. For each simulated trial, 2 internal risk models were developed on either the control population only (internal controls only) or the whole trial population blinded to treatment (internal whole trial). Bias was estimated for the internal models by comparing treatment effect predictions to predictions from the external model. Under all treatment assumptions, internal models introduced only modest bias compared with external models. The magnitude of these biases was slightly smaller for internal whole trial models than for internal controls only models. Internal whole trial models were also slightly less sensitive to bias introduced by overfitting and less sensitive to falsely identifying the existence of variability in treatment effect across the risk spectrum compared with internal controls only models.

Conclusions— Appropriately developed internal models produce relatively unbiased estimates of treatment effect across the spectrum of risk. When estimating treatment effect, internally developed risk models using both treatment arms should, in general, be preferred to models developed on the control population. (Circ Cardiovasc Qual Outcomes. 2014;7:00-00.)

Key Word: clinical trial ■ individualized medicine ■ risk

Randomized controlled trials are generally considered positive if the mean outcome in treated patients is superior to the mean outcome in the control arm. Although this approach has the virtue of simplicity, it has been criticized for offering limited guidance for whether individual patients ought to be treated. In some strongly positive trials, it is possible that a majority of patients receive little to no benefit, or even harm, from treatment. Unfortunately, patients have so many characteristics that might potentially influence the benefit of therapy, examining each using traditional subgroup analyses is often not informative. In addition to the well-appreciated issues of multiplicity leading to false-positives and low power leading to false-negatives, traditional subgroup analyses may fail to detect clinically relevant differences in treatment effects between groups that can be identified only when a combination of clinical variables is considered.

It is increasingly recognized that a patient’s baseline risk is a fundamental determinant of treatment benefit. Indeed, many treatment guidelines have adopted risk-sensitive treatment recommendations. Perhaps the most notable examples within the field of preventive cardiology are recommendations to initiate statin or aspirin therapy contingent on the baseline risk of cardiovascular disease. Although there are many examples in other fields as well, empirical evidence supporting these guidelines is not always clear, and details on how to implement guidelines is often lacking. A better evidence base for risk-based treatment recommendations might be available if clinical trials were analyzed using multivariable risk prediction methods, and it has recently been proposed that such methods be routinely applied. This approach can provide better estimates of how the degree of benefit or harm of a medical intervention varies across patients in the trial based on their overall risk of the study’s outcomes. This technique relies on a modeling approach that separately estimates baseline risk, typically, although not necessarily, using an independently derived predictive model to estimate the baseline outcome risk for individual study subjects (eg, Framingham risk score for the prediction of stroke/myocardial infarction risk), and then estimates treatment effect as a function of this baseline risk. Such a model tests whether lower-risk subjects receive similar

Received July 23, 2013; accepted December 6, 2013.
From the Departments of Neurology (J.F.B.) and Internal Medicine (R.A.H.), University of Michigan, Ann Arbor, MI; Department of Veterans Affairs, VA Center for Clinical Management and Research, Ann Arbor VA Healthcare System, Ann Arbor, MI (J.F.B., R.A.H.); and Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA (J.P.N., D.M.K.).


Correspondence to James F. Burke, MD, MS, Robert Wood Johnson Clinical Scholars Program, North Campus Research Center, 2800 Plymouth Rd, Building 10-G016, Ann Arbor, MI 48109. E-mail jamesbur@umich.edu

© 2014 American Heart Association, Inc.
Circ Cardiovasc Qual Outcomes is available at http://circoutcomes.ahajournals.org

DOI: 10.1161/CIRCOUTCOMES.113.000497
WHAT IS KNOWN

• For many treatments, there is known variability in the response to treatment.
• Even in some positive clinical trials, many patients receive little to no benefit, and some may even receive net harm.
• Identifying which patients are most likely to benefit from therapy is challenging.
• Multivariable risk prediction is a promising technique to determine whether treatment effect differs across patients (ie, whether heterogeneity of treatment effect exists), which relies on estimating treatment effect based on an individual’s baseline outcome risk. This technique requires a model to estimate a patient’s baseline risk.

WHAT THE STUDY ADDS

• In many cases, a valid external risk model may not exist for a given outcome, and it is tempting to develop such a model in a trial.
• In this study, we found that under a variety of treatment assumptions, internal risk models that are appropriately developed within a clinical trial introduce little bias compared with optimal external models.
• Internal models should have an important role in targeting cardiovascular treatments to appropriate treatment populations when external models are not available.

proportional benefit as higher-risk subjects using a treatment-baseline risk interaction term. Although not exploring every potentially clinically relevant hypothesis regarding heterogeneity of treatment effect (HTE), this approach examines the influence of a highly relevant mathematical determinant of treatment effect that integrates many patient characteristics in a multivariable framework, greatly mitigating some of the limitations of conventional subgroup analyses by minimizing the risk of multiple comparisons producing spurious findings,5,13 as well as delivers individualized estimates of treatment effect.14-17 An additional virtue of this approach is that even when the proportional treatment effect does not differ across the spectrum of baseline risk (ie, relative risk is constant across the baseline risk spectrum), this approach is well suited to showing how absolute treatment benefit varies based on baseline risk, which is often the most relevant metric for clinical decision making.

One proposal suggests that externally developed risk prediction tools be routinely used for multivariable risk assessment of HTE.10 However, some clinical trials lack validated pre-existing risk models for their main outcome (particularly composite outcomes), or the data may in other respects not be fully compatible with a published model (eg, model variables may be differently defined). In cases where no well-accepted external model exists to estimate baseline outcome risk, it is tempting to use the randomized controlled trial data to design a baseline risk model and examine treatment effect across risk defined by that internal risk model, an approach recently applied to the Justification for the Use of Statins in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER) study.17 Yet whether internally developed risk tools may introduce bias in the estimation of treatment effect across different risk categories has not been formally evaluated. An additional important question is what are the relative merits and risks of developing internal models on the whole trial population, including those patients who receive treatment, versus on the untreated (control) arm only. Deriving a model using only the control arm has intuitive appeal because arguably such a model best represents the patients’ pretreatment baseline risk, yet this approach may also lead to differences in model fitness between trial arms, thus inducing bias. Using simulation analyses based on cardiovascular disease prevention trials, we sought to examine whether using internal models, developed either on the whole trial or in the controls only, compared with using an external model, would result in biased estimates of the relationship between baseline risk and treatment effect.

Methods

Overview
To compare internal and external model bias and determine whether internal whole trial (IWT) or internal controls only (ICO) models lead to less bias and under what circumstances, we compared ICO, IWT, and external models under a variety of treatment effect scenarios. Specifically, we designed 4 different treatment effect scenarios that varied the magnitude of treatment effect and whether there was a main treatment effect (ie, average positive effect in the study population representing a constant relative risk reduction), risk-stratified HTE (ie, treatment effect varied as a function of a subject’s baseline risk for the study’s outcome), or both. For each of the 4 scenarios, we simulated a large study population with risk factors that predicted outcomes in a prespecified pattern (true risk factors), other potential risk factors that correlated with those risk factors and with outcomes (confounders), and prespecified treatment effects. Thereby, we were able to create an external model with optimal validity (ie, a model that would result from a large unbiased sample of the parent population) to estimate baseline risk for each scenario. Next, we repeatedly randomly sampled, with replacement, from each of the 4 parent populations to determine the statistical power and potential estimation biases in examining risk-stratified HTE in trials of varying sizes if we used an IWT model to predict baseline risk, versus an ICO model, versus an optimized valid external model.

We chose a simulation approach for this analysis instead of estimating these effects in an actual trial because of the intrinsic challenge of determining ground truth when applications of different models arrive at discordant conclusions. Given that this simulation approach requires some assumptions, we also applied a similar analytic approach to an actual trial where a widely used external risk model could be compared with an internal model. Specifically, we compared the Framingham risk score14 (external model) to internal models in the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT)10 to determine whether widely different interpretations may arise, thereby calling into question simulation assumptions (see the Data Supplement). The LRC-CPPT analysis plan was reviewed and approved by the Tufts Institutional Review Board. The authors had complete access to all LRC-CPPT data.

Simulated Treatment Scenarios/Patient Populations

Treatment scenarios were conceptually formulated as medication-based cardiovascular disease prevention trials. In all of the scenarios, patients were at risk for cardiovascular events, and the magnitude of this risk could be predicted based on the presence or
The 4 treatment scenarios were differentiated based on the prespecified treatment effects. Two different types of treatment effects were specified. First, there was a mean overall treatment effect—whether the treatment group had, on the whole, favorable outcomes compared with the control group. This effect represented the standard definition of whether a trial is positive. Second, we specified HTE—varying with the control group. This effect represented the standard definition of whether a trial is positive. Second, we specified HTE—varying with the control group. This effect represented the standard definition of whether a trial is positive.

### Trial Simulations

Individual trials, and their results, were simulated by repeated random draws with replacement from each scenario’s large parent population. In the base case, we drew 3000 intervention subjects and 3000 control subjects for each of the 4 scenarios. This sample size was selected for 2 reasons. First, it represented a well-powered study for detecting a main absolute outcome difference of 1.5% between the control and treatment arms. Second, it conformed to an oft-cited heuristic for adequate sample size to enable development of an ICO model without significant overfitting because it included ≈10 events per predictor variable (EPV) in the control arm. For each treatment scenario, we performed 1000 simulated trials and analyzed each trial for the treatment’s main effect and its interaction with baseline risk using treatment outcome models developed in each trial.

### Model Development

Before performing the simulated trials, we developed an optimized valid external risk model for the large parent population using only the 6 true risk factors in each of the treatment scenario populations using logistic regression. Results generated using this external model serve as the gold standard for comparison against results generated using the internal models. Then within each of the simulated trials, we used logistic regression to develop ICO and IWT baseline risk models predicting cardiovascular outcomes using both the 6 true risk factors as well as the 6 confounders as predictor variables. The ICO model was derived on the control trial population only, whereas the IWT model was derived on the control trial population only—both the control and treatment arms. For each subject in each simulated trial, we were then able to develop separate risk estimates from each of the 3 models: external, IWT, and ICO.

### Estimating Treated and Untreated Outcomes

Mean treatment effect and HTE were evaluated using logistic regression including 3 terms: predicted risk from the baseline risk model (IWT, ICO, external), a treatment indicator variable, and an interaction between treatment and baseline risk. The treatment indicator coefficient measures whether, and to what extent, a constant relative risk reduction exists across the risk spectrum, whereas the treatment–baseline risk interaction term measures whether differential treatment benefit exists on the part of the risk spectrum (eg, higher relative risk reduction in high- versus low-risk patients). These models enabled the estimation of an individual’s risk of the overall outcome with and without treatment for each of the 3 baseline risk models. For analyses that compare regression coefficients, baseline risk information was included in each treatment model using the percentile rank of risk from each baseline risk model to standardize baseline risks across models. Without such standardization, baseline risk would be systematically lower or higher for the IWT group compared with the external and ICO models depending on the direction of treatment effect.

### Estimating Model Bias and Statistical Power

Bias was estimated for each of the 5 trial scenarios using all 3 models using 3 separate approaches. First, the proportion of cases where the internal model outcome risk confidence interval included the optimized valid external point estimate was tabulated. Second, the average point estimate was estimated for each model type over each risk decile and compared with the estimated actual risk reduction (untreated risk–treated risk) across the series of simulated trials. Finally, the treatment and HTE regression coefficients and the degree of confidence interval coverage of the internal models were compared with the external model.

Power was estimated by the percentage of statistically significant main effects and interaction effects, when these effects were present in the parent population. Sensitivity, specificity, and intermodel reliability using Cohen κ were estimated by comparing the classifications of both coefficients of the ICO and IWT models to those using the optimized valid external model.

### Sensitivity to Overfitting

Our base case trial scenarios were designed so that the ratio of EPV included in the model was slightly >1:10 to limit bias due to overfitting. In real-world scenarios, it may be that there are a large number of potentially important predictors or a relatively small number of outcomes, and thus the models will be more susceptible to overfitting. To test how ICO and IWT models perform in these contexts, we set up a separate series of trials by varying the number of patients to adjust the EPV from 5 to 15.

### Results

#### Predictiveness of Baseline Risk Models

The predictiveness of the 3 baseline risk models is displayed in Table 2. As expected, the IWT and ICO models were slightly more predictive of outcomes in the simulated trials compared with the external model when <10 predictor variables per outcome were included. This increased predictiveness, however, is merely an indication of modest overfitting in the IWT and ICO models because the external model represents an optimized valid model, and predictiveness for both IWT and ICO models decreased as overfitting was reduced.

### Model Bias—Treatment Effect Comparisons Across Model Types

In the base case, which had 10 control EPVs, we found minimal bias in treatment outcome risk predictions when comparing the true treatment effects (those found using an optimized valid external model), compared with either internal model. The confidence intervals for point estimates from both internal models included the treatment outcome predicted by the external models for every treatment scenario in nearly all cases—99.7% of ICO and 99.9% of IWT outcome risk predictions included the external risk point estimate within their confidence intervals.

### Table 1. Description of Treatment Effects in 4 Different Treatment Scenarios

<table>
<thead>
<tr>
<th>HTE</th>
<th>No HTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean treatment effect</td>
<td>Scenario 1</td>
</tr>
<tr>
<td>No mean treatment effect</td>
<td>Scenario 2</td>
</tr>
</tbody>
</table>

HTE indicates heterogeneity of treatment effect.
Both ICO and IWT models produced, on average, minimal bias in the estimation of mean treatment effect compared with the optimal valid external model. The distribution of mean ICO biases was slightly narrower than the distribution of mean IWT biases, and these results for >1000 simulations are displayed in Figure 1.

The minimally biased predicted treatment effects of IWT and ICO models hold across the risk spectrum—neither IWT nor ICO resulted in substantial bias in either high- or low-risk patients regardless of treatment status or the presence/absence of HTE. IWT models predicted a fairly similar treatment effect size across the risk spectrum compared with external models. Conversely, ICO models tended to slightly underpredict treatment effect in lower-risk patients and overpredict treatment effect size in the highest-risk patients. This effect is because of the model fit being better in the control group, potentially inducing a spurious risk-by-treatment interaction. However, the magnitude of both effects is modest in our simulations (Figure 2). Bias was further quantified by comparing the regression treatment–risk interaction and treatment indicator coefficients from the internal treatment outcome models to the external models, and the results are summarized in the Data Supplement.

Model Power
Because the overall treatment effect and HTE were specified in each of our treatment scenarios, we were able to determine how often the 3 modeling approaches correctly identified significant treatment indicator variables and treatment–interaction effects in >1000 simulated trials. Both IWT and ICO models had excellent agreement with external models on the treatment indicator variables (all k values >0.80). Conversely, ICO models had higher sensitivity (ie, a lower false-negative rate) but at a cost of a higher false-positive rate (ie, finding a significant effect when there was no HTE). For example, the false-negative rate in scenario 1, in which HTE was present, was only 25.3% by ICO versus 47.3% by IWT, but in scenario 3, in which there was no HTE, ICO models had a false-positive HTE of 14% compared with 5.4% for IWT models (Table 3).

Sensitivity to Overfitting
Risk of overfitting is an intrinsic weakness of ICO models compared with IWT models, given that they are derived on a smaller population. Potentially more important, differential fitting between treatment arms might induce or exaggerate HTE compared with an IWT model developed blinded to treatment. To explore the magnitude of these effects, we built IWT and ICO models with varying degrees of overfitting, ranging from highly overfit models (5 EPVs) to relatively conservative models (15 EPVs). For both models, the treatment effect bias was more tightly distributed around zero as the EPV increased. For all scenarios and for a given EPV, there was a slightly wider distribution of bias in ICO models than in IWT models (Figure 3). As anticipated, the direction of this bias tended to exaggerate HTE, with more benefit seen in higher-risk patients because of increased risk heterogeneity in the control arm.

Discussion
This study has 2 key findings. First, we found that when internal models are developed appropriately (having ≥10 control

### Table 2. Model Predictiveness

<table>
<thead>
<tr>
<th>Outcomes per E predictor</th>
<th>IWT (90% CI)</th>
<th>ICO (90% CI)</th>
<th>External (90% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean treatment effect present</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1: Mean treatment effect plus HTE</td>
<td>0.75 (0.70–0.79)</td>
<td>0.74 (0.71–0.77)</td>
<td>0.74 (0.71–0.76)</td>
</tr>
<tr>
<td>Scenario 2: Mean treatment effect plus no HTE</td>
<td>0.73 (0.69–0.77)</td>
<td>0.72 (0.69–0.75)</td>
<td>0.72 (0.69–0.74)</td>
</tr>
<tr>
<td>Scenario 3: No mean treatment effect plus HTE</td>
<td>0.75 (0.71–0.79)</td>
<td>0.74 (0.71–0.77)</td>
<td>0.74 (0.71–0.76)</td>
</tr>
<tr>
<td>Scenario 4: No mean treatment effect plus no HTE</td>
<td>0.73 (0.69–0.77)</td>
<td>0.73 (0.70–0.75)</td>
<td>0.72 (0.69–0.76)</td>
</tr>
</tbody>
</table>

For each of the 3 models, the predictiveness of the baseline risk prediction model (predicting cardiovascular outcomes in the trial) is displayed. The 90th percentile CIs are displayed in parentheses for all models except for the external risk prediction models because there was only 1 external risk prediction model for each treatment scenario. CI indicates confidence interval; ICO, internal controls only; and IWT, internal whole trial.

**Figure 1.** Mean bias in treatment effect. Box plot representing the distribution of mean treatment effect bias (difference in predicted absolute risk reduction [ARR]) from the internal whole trial (IWT) and internal controls only (ICO) models compared with the external model. The box represents the 25th, 50th, and 75th percentiles, and the whiskers represent the 5th and 95th percentiles.

**Figure 3.** Difference in ARR Compared to External Model. Diagram showing the distribution of ARR from the IWT and ICO models compared with the external model for each scenario. The box represents the 25th, 50th, and 75th percentiles, and the whiskers represent the 5th and 95th percentiles.
group EPVs), they produced relatively unbiased estimates of
treatment effect across the spectrum of baseline risk. Second,
we found that IWT models should, in general, be preferred to
ICO models as they offer narrower distribution of treatment
effect biases, slightly less sensitivity to overfitting, and less
susceptibility to false-positive identification of HTE.

Under a range of treatment effect scenarios, both ICO and
IWT models arrived at similar estimates of the study popula-
tion’s heterogeneity of baseline risk and how the relative and
absolute treatment effect varied as a function of baseline risk
(risk-based HTE), just as long as model development consid-
ered no more than 1 risk factor for every 10 outcomes present

Table 3. Power Estimates

<table>
<thead>
<tr>
<th>Treatment Scenario</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>External model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment coefficient</td>
<td>Significant N</td>
<td>312</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td>HTE coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IWT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTE coefficient</td>
<td>Significant N</td>
<td>529</td>
<td>494</td>
<td>55</td>
</tr>
<tr>
<td>Sensitivity vs external</td>
<td>89.04%</td>
<td>85.83%</td>
<td>63.64%</td>
<td>54.76%</td>
</tr>
<tr>
<td>Specificity vs external</td>
<td>88.32%</td>
<td>86.76%</td>
<td>97.99%</td>
<td>94.98%</td>
</tr>
<tr>
<td>κ vs external</td>
<td>0.7733</td>
<td>0.7259</td>
<td>0.6216</td>
<td>0.374</td>
</tr>
<tr>
<td>Internal model: ICO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTE coefficient</td>
<td>Significant N</td>
<td>747</td>
<td>733</td>
<td>140</td>
</tr>
<tr>
<td>Sensitivity vs external</td>
<td>98.87%</td>
<td>97.98%</td>
<td>61.82%</td>
<td>52.38%</td>
</tr>
<tr>
<td>Specificity vs external</td>
<td>52.44%</td>
<td>50.79%</td>
<td>88.78%</td>
<td>88.92%</td>
</tr>
<tr>
<td>κ vs external</td>
<td>0.5264</td>
<td>0.4849</td>
<td>0.2929</td>
<td>0.2087</td>
</tr>
</tbody>
</table>

For all 3 models, the number of trials (out of 1000) in which the treatment and HTE coefficients were statistically significant. For the ICO and
IWT models, the categorization of significant coefficients is compared with external models using sensitivity, specificity, and κ. HTE indicates
heterogeneity of treatment effect; ICO, internal controls only; and IWT, internal whole trial.
in the control group (10 EPVs). This implies that the development of internal models is a practical strategy to test for HTE. This has the potential to greatly increase the degree to which clinical trial results can be used to better inform individual patient treatment decisions in instances in which external risk models do not already exist for a study’s outcome measure. We would still argue that when valid externally developed prediction tools are available, their use should be preferred, not because of greater internal validity but because the use of an external tool is a better test of the external validity of HTE results, likely offers superior calibration, and permits easier translation into practice through a variety of techniques.

In particular, to make risk-based treatment decisions, the risk model must be well calibrated to the target population. Given the differences between patients that enroll in trials and patients in the broader population, this is an important limitation of directly applying internal risk models to the broader target population. Consequently, if clinically important HTE is discovered with an internally developed model, it should motivate the more difficult step of developing and validating an appropriate risk model for clinical translation.

When developing internal models, this study suggests that IWT models be favored over ICO models. This finding was mainly because of IWT models being insensitive to overfitting, whereas overfitting ICO models differentially across study arms leads to more bias in the magnitude of risk-based HTE. Furthermore, even in the absence of overfitting, ICO models are intrinsically more susceptible to falsely identifying risk-based HTE when none exists because of an underestimation of the true standard error.

In addition, this study illustrates the importance of avoiding overfitting by ensuring an adequate number of events per variable in the baseline risk model. We observed an increase in bias in both IWT and ICO models as EPV decreased. Our data suggest that applying the conventional rule of thumb (≥10 predictor variables for each binary outcome in the data set) results in the introduction of minimal bias into treatment effect estimates across the spectrum of baseline risk, although this should ideally be tested in other scenarios. Although this might create a problem in some small studies or studies with a small number of outcomes, for most clinical trials, following this rule of thumb will merely require only considering risk factors for which there is strong empirical or theoretical reasons to suspect that they are substantially associated with the risk of the outcome.

This study is limited by the assumptions of our simulation environment. Although this study explored a wide variety of treatment effects, in order to maintain a comprehensible degree of complexity, we made some potentially limiting assumptions in setting up our treatment scenarios. Most notably, we modeled a disease–treatment context with relatively low outcome risk, relying on moderately predictive models in the context of moderate degrees of confounding and assuming a constant risk of baseline adverse effects. Although assuming a constant absolute risk of adverse effects across patients at different outcome risks may seem like a substantive limitation, it is a commonly applied assumption and studies that have modeled differential risks of benefits and harms have not found substantial divergence. Regardless, when modeling actual trials, the possibility of differential harms across risk strata should be explicitly explored empirically. Although it is uncertain whether our conclusions would hold under differing assumptions, these parameters were chosen because they were thought to be applicable to a wide variety of real-world clinical contexts.

Although simulation studies require potentially limiting assumptions, those assumptions were necessary in this context because of the intrinsic challenges of studying this question using data from actual trials. If internal and external models applied to an actual trial were discordant, it would be impossible to determine whether that disagreement reflected limited calibration of the external model in the trial population or biases in the internal model. Thus, this study provides important evidence that using carefully developed internal models results in similar estimates of treatment effect size and similar estimates of how treatment effect varies as an effect of risk compared with optimized valid external models.
Internal models are powerful tools to estimate treatment effect across the risk spectrum. Because these models seem to introduce minimal bias compared with external models, they should have an important role in targeting cardiovascular treatments to appropriate treatment populations when external models are not available.

Acknowledgments
This article was prepared using LRC-CPTT research materials obtained from the National Heart, Lung, and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the LRC-CPTT or the National Heart, Lung, and Blood Institute. We also thank Tim Hofer for his assistance in developing the code to develop the simulated populations.

Sources of Funding
Funded by a Patient Centered Outcome Research Institute Methodology Grant. J.F.B. was funded by an Advanced Fellowship from the Department of Veteran’s Affairs and by National Institute of Neurological Disorders and Stroke grant number 1K08NS082597.

Disclosures
None.

References
Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials
James F. Burke, Rodney A. Hayward, Jason P. Nelson and David M. Kent

Circ Cardiovasc Qual Outcomes. published online January 14, 2014;
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2014 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/early/2014/01/13/CIRCOUTCOMES.113.000497

Data Supplement (unedited) at:
http://circoutcomes.ahajournals.org/content/suppl/2014/02/12/CIRCOUTCOMES.113.000497.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org//subscriptions/
Supplemental Material

LRC-CPPT Analysis

To explore the possibility that the findings from our primary simulation analysis may be susceptible to bias due to assumptions in the simulation environment, we explored applying similar analytic techniques to the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) trial. LRC-CPPT compared cholestyramine to placebo for the reduction of cardiovascular events in men with hyperlipidemia. LRC-CPPT was selected because of its large sample size and because the trial data was collected and reported in the trial dataset in such a way that Framingham risk scores could be directly estimated. Thus, we were able to compare internal models developed in the trial to a widely used external model, the Framingham risk score. In LRC-CPPT, the main results were: XX and we are unaware of prior reporting of HTE.

Methods

Logistic regression was used to estimate baseline internal whole trial (IWT) and internal controls only (ICO) risk models using age, race, smoking status, total cholesterol, high density lipoprotein (HDL) and systolic blood pressure as predictor variables. The IWT model was developed on the whole trial dataset and the ICO model on the controls only.

Mean treatment effect and heterogeneity of treatment effect (HTE) were estimated for each of the baseline risk models using logistic regression and including baseline risk, treatment status and a risk-treatment interaction term. This process was repeated over 1000 bootstrapped samples to generate estimates of uncertainty in ARR.

Results

All three models were only moderately predictive with c-statistics of 0.59 (external) vs. 0.66 (IWT) vs. 0.64 (ICO). Thus, these models were somewhat less predictive than our simulated examples and, unlike our simulated examples the IWT model was slightly more predictive than the ICO model. Consistent with the reported trial analysis, we found a mean treatment effect of borderline significance in all three models and did not find HTE as all treatment status coefficients and treatment-risk interaction terms were non-significant.

There was little bias in estimates of treatment effect across the risk spectrum in CPPT for either internal model. In the highest risk deciles, there was slightly less bias, compared to the external model, for the IWT model vs. the ICO model. (Figure)
**Figure:** The outcome/risk distribution for each of three types of models in CPPT. Estimated treatment effects are on the y-axis and the decile of baseline risk on the X-axis. Error bars signify 95% confidence intervals over bootstrapped samples.
Both internal models tended to slightly overestimate the treatment indicator coefficient and under-estimate the treatment-interaction coefficient across the set of treatment scenarios. For all five scenarios, the magnitude of the coefficient bias was slightly larger for the ICO models than for the IWT models. IWT models also tended to be more conservative than ICO models with a larger ratio standard error/coefficient ratio than IWT models for both treatment and treatment-risk interaction coefficients. As a consequence of both the lesser degree of mean coefficient bias and the larger standard error/coefficient ratios, IWT models consistently covered a larger proportion of the coefficient confidence intervals relative to external models. (e-Table) These differences in coefficients lead to small biases in treatment risk predictions.

<table>
<thead>
<tr>
<th>Scenario 1: Mean Treatment effect + HTE</th>
<th>IWT mean (90% CI)</th>
<th>ICO mean (90% CI)</th>
<th>External mean (90% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Coefficient</td>
<td>0.99 (-0.31 - 2.42)</td>
<td>1.57 (0.27 - 3.04)</td>
<td>0.74 (-0.45 - 1.99)</td>
</tr>
<tr>
<td>Treatment Coefficient Bias</td>
<td>0.25 (-0.29 - 0.90)</td>
<td>0.83 (0.18 - 1.73)</td>
<td></td>
</tr>
<tr>
<td>Treatment Coefficient Standard Error</td>
<td>0.77 (0.63 - 0.98)</td>
<td>0.77 (0.63 - 0.99)</td>
<td>0.69 (0.58 - 0.84)</td>
</tr>
<tr>
<td>Treatment Confidence Interval Coverage</td>
<td>92.3% (75.0% - 100.0%)</td>
<td>74.6% (42.6% - 100.0%)</td>
<td></td>
</tr>
<tr>
<td>Treatment-Risk Interaction Coefficient</td>
<td>-0.017 (-0.035 - 0.000)</td>
<td>-0.026 (-0.045 - -0.008)</td>
<td>-0.014 (-0.030 - 0.002)</td>
</tr>
<tr>
<td>Treatment-Risk Coefficient Bias</td>
<td>-0.003 (-0.012 - 0.005)</td>
<td>-0.012 (-0.025 - -0.002)</td>
<td></td>
</tr>
<tr>
<td>Treatment-Risk Coefficient Standard Error</td>
<td>0.010 (0.009 - 0.012)</td>
<td>0.010 (0.009 - 0.013)</td>
<td>0.010 (0.008 - 0.011)</td>
</tr>
<tr>
<td>Treatment-Risk Confidence Interval Coverage</td>
<td>91.4% (73.5% - 100.0%)</td>
<td>72.4% (36.5% - 100.0%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 2: No mean effect + HTE</th>
<th>IWT mean (90% CI)</th>
<th>ICO mean (90% CI)</th>
<th>External mean (90% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Coefficient</td>
<td>1.59 (-0.06 - 3.50)</td>
<td>2.25 (0.72 - 4.26)</td>
<td>1.12 (-0.22 - 2.80)</td>
</tr>
<tr>
<td>Treatment Coefficient Bias</td>
<td>0.47 (-0.21 - 1.37)</td>
<td>1.13 (0.29 - 2.40)</td>
<td></td>
</tr>
<tr>
<td>Treatment Coefficient Standard Error</td>
<td>0.93 (0.71 - 1.26)</td>
<td>0.95 (0.72 - 1.33)</td>
<td>0.80 (0.63 - 1.06)</td>
</tr>
<tr>
<td>Treatment Confidence Interval Coverage</td>
<td>90.4% (69.6% - 100.0%)</td>
<td>71.6% (35.8% - 100.0%)</td>
<td></td>
</tr>
<tr>
<td>Treatment-Risk Interaction Coefficient</td>
<td>-0.023 (-0.048 - -0.001)</td>
<td>-0.033 (-0.057 - -0.013)</td>
<td>-0.017 (-0.039 - 0.002)</td>
</tr>
<tr>
<td>Treatment-Risk Coefficient Bias</td>
<td>-0.006 (-0.018 - 0.004)</td>
<td>-0.016 (-0.033 - -0.003)</td>
<td></td>
</tr>
<tr>
<td>Treatment-Risk Coefficient Standard Error</td>
<td>0.012 (0.010 - 0.016)</td>
<td>0.013 (0.010 - 0.016)</td>
<td>0.011 (0.009 - 0.014)</td>
</tr>
</tbody>
</table>
### Table 3: Coefficient-level Bias

For all treatment scenarios and all three predictive models, the logistic regression coefficients, standard errors for those coefficients and bias relative to the external model is displayed for both the treatment coefficient and the treatment-risk interaction coefficients. In addition, the proportion of the coefficient confidence interval, defined by the external model, is displayed for both ICO/IWT models.

<table>
<thead>
<tr>
<th>Scenario 3: No effect or HTE</th>
<th>Treatment Coefficient</th>
<th>Treatment Coefficient Bias</th>
<th>Treatment Coefficient Standard Error</th>
<th>Treatment Confidence Interval Coverage</th>
<th>Treatment-Risk Interaction Coefficient</th>
<th>Treatment-Risk Coefficient Bias</th>
<th>Treatment-Risk Coefficient Standard Error</th>
<th>Treatment-Risk Confidence Interval Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Coefficient</td>
<td>-0.14 (-2.07 - 1.71)</td>
<td>1.26 (-0.38 - 3.05)</td>
<td>0.03 (-0.96 - 0.88)</td>
<td>93.8% (78.2% - 100.0%)</td>
<td>0.002 (-0.023 - 0.026)</td>
<td>-0.018 (-0.040 - 0.003)</td>
<td>0.000 (-0.011 - 0.013)</td>
<td>92.1% (75.1% - 100.0%)</td>
</tr>
<tr>
<td>Treatment Coefficient Bias</td>
<td>1.01 (0.79 - 1.31)</td>
<td>0.96 (0.74 - 1.26)</td>
<td>0.013 (0.011 - 0.016)</td>
<td>69.2% (29.8% - 100.0%)</td>
<td>-0.018 (-0.040 - 0.003)</td>
<td>0.001 (-0.020 - 0.023)</td>
<td>0.000 (-0.011 - 0.013)</td>
<td>62.9% (15.9% - 100.0%)</td>
</tr>
<tr>
<td>Treatment Confidence Interval Coverage</td>
<td>93.8% (78.2% - 100.0%)</td>
<td>67.2% (25.4% - 100.0%)</td>
<td>72.8% (34.0% - 100.0%)</td>
<td>93.5% (74.9% - 100.0%)</td>
<td>0.002 (-0.023 - 0.026)</td>
<td>-0.018 (-0.040 - 0.003)</td>
<td>0.001 (-0.011 - 0.013)</td>
<td>62.9% (15.9% - 100.0%)</td>
</tr>
</tbody>
</table>

### Scenario 4: Mean effect + no HTE

| Treatment Coefficient | -1.15 (-4.73 - 1.53) | 0.70 (-1.83 - 3.07)       | -0.07 (-1.94 - 1.44)                  | 1.54 (1.00 - 2.58)                   | 1.78 (0.14 - 4.18)               | 0.88 (1.82 - 1.24)         | 0.00 (0.011 - 0.016)                 | 91.8% (72.4% - 100.0%)                    |
| Treatment Coefficient Bias | 1.54 (1.00 - 2.58) | 0.88 (1.82 - 1.24)        | 0.001 (-0.019 - 0.023)                | 0.019 (0.014 - 0.030)                | -0.025 (-0.057 - 0.002)           | 0.017 (0.013 - 0.022)        | 0.016 (0.012 - 0.024)                 | 67.9% (24.4% - 100.0%)                    |
| Treatment Confidence Interval Coverage | 91.8% (72.4% - 100.0%) | 67.9% (24.4% - 100.0%) | 72.8% (34.0% - 100.0%)               | 93.5% (74.9% - 100.0%)               | 0.004 (-0.032 - 0.048)              | -0.021 (-0.053 - 0.010)        | 0.001 (-0.019 - 0.023)                 | 67.9% (24.4% - 100.0%)                    |

**e-Table 3:** Coefficient-level Bias.