Background and Purpose—The Carotid Revascularization Endarterectomy Versus Stenting Trial was completed with a low stroke and death rate. A lead-in series of patients receiving carotid artery stenting was used to select the physician-operators for the study, where performance was evaluated by complication rates and by peer review of cases. Herein, we assess the potential contribution of statistical evaluation of complication rates.

Methods—The ability to discriminate between stent operators who can successfully meet the published guideline of <3% combined rate of stroke and death is calculated under the binomial distribution, based on a small consecutive case series (n=24 patients).

Results—A criterion of ≤2 stroke or death events among the 24 patients (<8% event rate) was required of operators. Setting such a high criterion, however, ensures an inability to exclude operators who cannot meet the criteria. In fact, if a good operator is defined as having a 2% event rate and a poor operator as a 6% event rate, even a series of 240 patients would (on average) still exclude 5.4% of the good operators and include 4.6% of the poor operators.

Conclusions—The low periprocedural event rates in the trial suggest success in separating skillful operators from less skillful. However, it seems unlikely that statistical assessment of event rates in the lead-in contributed to successful selection, but rather successful selection was more likely because of peer review of subjective and other factors including patient volume and technical approaches.

Clinical Trial Registration—URL: http://www.clinicaltrials.gov. Unique identifier: NCT00004732.

(Stroke. 2014;45:3325-3329.)

Key Words: clinical trial ■ quality control ■ stroke

Although carotid endarterectomy was an established procedure, CAS was a newer procedure when CREST was initiated, and so the investigators implemented a rigorous assessment of potential interventionists, including in a lead-in registry for stent operators.5,6 During this process, the interventional management committee reviewed 10,164 CAS cases from 427 interventionalist-applicants (an average of 24 from each operator). The feasible size of the series of patients conducted before the admission of the operator to the study is a serious consideration. Even for an active operator conducting a procedure a week, the stringent requirements of CREST required that ≈6 months work need be completed for an operator to be admitted to the study. As such, increases in the size of the qualifying series substantially beyond the neighborhood of 24 are practically infeasible. Accordingly, several criteria were used to approve operators, which have been previously described as “a low complication rate, experience with ≥15 procedures, use of proper standard carotid stenting technique,
and avoidance of erroneous techniques (eg, improper device use, inappropriate balloon sizing, use of 0.035" wires, use of general anesthesia). They can be classified as evaluating the performance (the low complication rate), establishing the experience (≥15 procedures), and the reviewing of technique (use of proper technique and avoidance of erroneous techniques). In CREST, of the 427 applicants for participation as an operator, 224 (52%) were approved for randomization; among those 427 applicants, 238 were required to perform cases in the lead-in and 158 of those applicants were approved for randomization, based on the lead-in data. The success of the study in identifying operators with low event rates stands as proof that the process worked, but can this success be attributed to exclusions on the basis performance, experience, or technique? Herein, we review the contribution to the statistical assessment of performance in this lead-in registry and the likelihood that it made a meaningful contribution to selecting the operators who can perform the procedure with a low risk of complication.

The focus of much of this report is on the ability to select operators based on the result of a series of 24 patients, reflecting the average number of patients available for decision making in the CREST trial. However, we also provide estimates of how large an increase in the available series of patients would be required to provide a reasonable basis for the identification of good operators and the exclusion of bad operators. This consideration of the review process is important more generally, as referring physicians sometimes take a similar approach to ensure the quality of care given to their patients, where requests for performance in recent patients is sometimes used to determine where to make referrals. In addition, patients are advised to seek information on qualifications of potential surgeons or operators before scheduling any nonemergency procedures, with lay magazines, such as Consumer Reports, recommending that patients ask 4 questions in selecting a surgeon before a procedure, including “What are your success, failure, and complication rates?” How, and when, should we look at the batting averages of successful procedures to select a surgeon, an operator, or for that matter ... a car mechanic?

### Methods

Statistically, assessing the performance of an individual surgeon or operator is guessing (statisticians prefer estimating) the true rate of poor outcomes after their procedures. Clearly, there are characteristics of the individual patient that could influence the rate of poor outcomes. For example, periprocedural event rates in CAS could be generally higher in the elderly population or in women. However, we assume that the goal is to assess the likelihood of stroke or death for an average patient for a specific operator, where a low average risk of events is likely the criteria for approving an operator (assumed to not change over time or over patients). For example, a low average risk of events is likely the criteria for approving an operator (assumed to not change over time or over patients).

One then can assume that there is a true event rate for each operator (assumed to not change over time or over patients). For example, the current guidelines suggest CAS and carotid endarterectomy be considered as treatment alternatives for severe asymptomatic carotid artery disease if they can be performed with an event rate <3% in asymptomatic patients. However, if there is variation among operators, how can a study, such as CREST, systematically choose study operators with event rates below this level? More generally, how can a patient be assured that the true event rate for the specific operator is below this level? One might assume that we could look at the observed event rates for a specific operator, and if he/she is a good apple, their event rate would be <3%, whereas if they are a bad apple, their event rate would be >3%. Unfortunately, the situation is more complex than this approach suggests.

For simplicity, assume that the patients treated by a specific operator has a chance of p of having a periprocedural stroke or death (ie, poor outcome or complication). Then we are interested in including the physicians with p <0.03 (the good apples) and excluding the physi-cians with p >0.03 (the bad apples). We can estimate an individual operator's p (we will refer to this estimate as b) by simply dividing the number of patients with a poor outcome by the number of patients studied (p = k/n, where k is the number with poor outcome and n is the number of patients studied).

There are several challenges to this approach. First, even without changes to the true chance of a poor outcome, chance may imply a different number of complications in 2 series of patients. For example, a single operator may perform 24 procedures with no complications, and then on the subsequent series of 24 patients, the same operator may have 3 poor outcomes. That is, the estimated event rate is only a reflection of the true event rate, and having an estimated event rate lower than 3% does not assure the reviewer that the true rate is acceptable. The second challenge is inherent in our goal to select operators only if they had zero (0) events, this would result in the exclusion of 38% of the good operators (the 2% column) and the inclusion of 3% of good operators (with a true 2% event rate) excluded plus the percent of poor operators (with a true 6%, 8%, or 10% event rate) included. For any sample size, we selected the threshold to minimize this error rate. For example, in Table 1, if the decision was to include operators only if they had zero (0) events, this would result in the exclusion of 38% of the good operators (the 2% column) and the inclusion of 23% of the poor operators (the 6% column) for an error rate of 61%. This error rate is smaller than including operators with ≤1 events (an error rate of 66%) or to include operators with ≤2 events (an error rate of 84%). Therefore, we would declare the optimal rule for a series of 24 patients to be to include operators with no events, with an unfortunately high error rate of 61%. We then repeated this process for sample size ranging up to 500 and plotted the decline in error rate with the increasing size of the series. We suggest that to make reliable decisions, it would be optimal to exclude only ≤5% of good operators and include no >5% of poor operators, so a goal of an error rate of <10% would seem reasonable.

### Results

#### Approaches to Keep the Good Apples

Suppose that a particular operator is a good apple, that is his/her true complication rate is ≤3%. However, even among the good operators, there are those who are outstanding (true
Table 1. Distribution of Operators on the Anticipated Number of Poor Outcomes as a Function of Their True Chance of Having a Poor Outcome on a Particular Patient

<table>
<thead>
<tr>
<th>No. of Poor Outcomes</th>
<th>True Chance of a Poor Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acceptable Operators (Good Apples)</td>
</tr>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>0</td>
<td>79</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6+</td>
<td>0</td>
</tr>
</tbody>
</table>

Of 100 operators with a specific true chance of a bad outcome, the anticipated number with a particular number of bad outcomes. For example, if there were 100 operators with a 3% true chance of a bad outcome, we would anticipate 48 operators to have no bad outcomes, 36 to have 1 bad outcome, 13 to have 2 bad outcomes, and 3 to have 3 bad outcomes.

Event rate of 1% or \( p=0.01 \), those who are good (true event rate of 2% or \( p=0.02 \)), and those who just barely meet the criteria (true event rate of 3% or \( p=0.03 \)). Then suppose we require each operator to perform the procedure on a series of 24 patients. Then out of 100 potential outstanding operators who each perform 24 procedures, on average how many operators will have no complications, how many 1 complication, how many 2 complications, and so forth? Similarly, what would the outcome be for 100 good operators or 100 who just meet the criteria? For each of these good type of operators, Table 1 shows the expected number of operators who have no events, 1 event, 2 events, 3 events, 4 events, 5 events, or ≥6 events. As can be seen, if we set the criteria for inclusion of an operator in the study as having 0 or 1 events, then we will only inappropriately exclude 2 of the outstanding operators; however, we will inappropriately exclude 8 of the good operators and inappropriately exclude 16 of the operators who barely meet inclusion criteria. As such, it would seem to meet the needs of the study that operators with ≤2 events be included in the study (inappropriately excluding no outstanding operators, a single good operator, and only 3 operators just meeting criteria). There seems to be some success in including operators meeting this criterion; however, to reliably do so, we would need to accept operators with event rates of ≤2 among 24 patients, an event rate of 8.3% or lower, not our ideal of <3%.

Approaches to Discard the Bad Apples

From a patient safety perspective, an even more important goal than including the operators meeting criteria (keeping good apples) is excluding the operators who fail to meet criteria (throwing away bad apples). Like there are a spectrum of abilities among the good operators, there are likely operators who barely fail to meet criteria (true event rate of 4% or \( p=0.04 \)), those that are moderately unacceptable (say a true event rate of 8% or even 10%, \( p=0.08 \) or \( p=0.10 \)), and those that are fundamentally unacceptable (say a true event rate of 20% or \( p=0.20 \)). Assuming that we have 100 of each of these types of operators, the expected number will have a range of events as shown in Table 1. For those barely not meeting the criteria (4% complication rate), if we set the criteria as above to be having ≤2 events, then we would include 94 out of 100 operators that we wish to exclude. Perhaps more disturbing, we would expect to include 70 out of 100 operators with a true 8% event rate (more than twice the practice guideline-defined limit) and 56 of 100 operators with a true 10% event rate (≥3× the limit). In fact, the exclusion of operators with high event rates can only be achieved if the true event rate is extraordinarily high, such as in the neighborhood of a true 20% complication rate. Even if we set the rate to having no complications, we would still include 23% of operators with a 6% event rate (twice the limit), 14% of operators with an 8% complication rate, and 8% of operators with a 10% complication rate. These data suggest that there is little hope of reliably excluding the bad apples from the study (unless they are truly unacceptable), and even setting the goal of no events among 24 cases will tend to include as many as 23% of those with twice the acceptable event rate.

Assessment of the Required Size of the Series to Make Reliable Decisions

Figure shows the error rate as a function of the number of patients in the series (using the optimal threshold for inclusion of an operator). It would require a series of 240 patients to have an error rate below 10% to distinguish between good operators with a 2% complication rate and poor operators with a 6% complication rate. With 240 patients, the optimal rule would be to include operators with ≤8 (3.3%) complications, and at this threshold, 5.4% of good operators would be excluded and 4.6% of poor operators would be included (error rate of 10%). If the goal was to distinguish between good operators with a 2% complication rate and poor operators with an 8% complication rate, a series of 120 patients would achieve a 10% error rate. To distinguish between operators with a 2% versus 10% event rate, 83 patients would achieve a 10% error rate.
Discussion

By peer review of a prospective consecutive case series averaging 24 patients per stent operator, CREST was shown to successfully choose operators with low complication rates. This process ultimately excluded 48% of operators who had applied to be part of the study. One would presume that the reason for the overall low event rate for the study was that these excluded operators included many bad apples that would not truly meet the criteria. Likewise, one would presume that the 52% who were accepted for the study included many good apples who do meet the criteria. Hence, the randomized trial phase of CREST suggests a proven ability to discriminate between the good and bad apples.

However, looking at complication rates in a series of this size could have provided only modest guidance to select those with good event rates. Recalling that the published criteria for selection were performance (ie, periprocedural stroke and death rate), volume, and peer review of technique, this would suggest that either the assessment of volume or the peer review of technique must have been the primary drivers of the success of the study. In their review of individual cases, the Intervventional Management Committee noted a diverse array of red flags, indicating that specific operators should be excluded, such as an administration of 12,000 IU Heparin, needing to use complex sheath access techniques or guiding catheters, using undersized stent diameters or excessive use of long stents, postdilatation with >5 mm balloon or doing multiple inflations, failing to urgently lower blood pressure (if it does not fall) after stent placement or an protection device time >15 minutes. Identifying these and other potential red flags requires the judgment of individuals with appropriate high levels of experience and expertise regarding the quality of the operator and the procedure, and these skills not easily quantified. As such, the study is grateful for what must be the skillful subjective review performed by the Interventional Management Committee (members shown in Table 2).

The central (statistical) problem in the assessment of performance is that it is difficult to reliably detect differences in binomial outcomes (in this case, in the neighborhood of p = 0.03). We demonstrate that to achieve an acceptably low error rate (sum of exclusion of good operators plus inclusion of poor operators), a series of 240 patients would be required. For an active operator performing a procedure a week, this would require >4.5 years to qualify for the study. Even with the entire CREST experience of 594 CAS procedures in asymptomatic patients and the lowest reported event stroke and death rate in any randomized trial (2.5%), the 95% confidence intervals for the event rate still extend up to 4.1%. Hence even with a series of 594 patients, we cannot definitively state that guideline level of 3.0% was met. This underscores not only the challenges of selecting operators for the CREST trial, but also the challenge of following the advice of Consumer Reports in the selection of surgeons to ask What is your success, failure, and complication rates?

A further complication is that patients may not be assigned to surgeons/operators in a random manner, but rather the more challenging patients are likely assigned to the better surgeon/operators. General approaches for adjustment for case-mix have been suggested; however, general rules of thumb require 10 to 20 events per variable considered for case-mix adjustment. Even with 594 asymptomatic patients in CREST treated with CAS, there were only 15 periprocedural events, making modeling for risk adjustment practically infeasible. Basically, the inability for case-mix modeling arises from the same basic issue as the challenge of statistical assessment of event rates … there is not a sufficient number of events for reliable statistical analysis.

Finally, the IMC in CREST was charged with selection of outstanding operators to ensure that stenting would be safely done in the randomization phase of the trial. It should be noted that minimization of risk from a procedure can be achieved by means beyond selection of skilled operators. There is growing interest and experience with the use of simulation as a means of reducing procedural risk. Additionally, although attention usually focuses on individual apples, one should not ignore the importance of the apple tree, ie, the environment in which the procedure is performed. For example, it is now considered standard of care for the cardiac operating room to have ongoing quality improvement projects, checklists, briefings, and formal handoff protocols. Presumably, these elements of care and process would also be of benefit in the setting of CAS.

In conclusion, CREST seems to have selected appropriate and skillful operators by a combination of reviewing performance, volume, and a subjective review of technique. Herein, we suggest that although there is some information to be gained in the statistical review of performance, the criteria for selection of the operator is so stringent that the information from a small case series is limited. It is uncomfortable for the statistician coauthors to admit that because the CREST review was successful, the greater value in the review would seem to have been from the subjective review of techniques or other sources of information (such as reputation). Statistics is a critically powerful tool to address many issues; however, unless large case series can be constructed, we suggest that it may not be the best tool for evaluating the ability of individual operators to meet stringent (<3%) performance rates, and results from small series (eg, ≤24) be interpreted with caution.
Sometimes you have to pick the good apples and throw away the bad by just looking at them instead of using a statistical assessment.

**Sources of Funding**

The Carotid Revascularization Endarterectomy Versus Stenting Trial was funded by the National Institute of Neurological Disorders and Stroke (US) [R01 NS038384] with additional support from Abbott Vascular, Inc.

**Disclosures**

None.

**References**

Picking the Good Apples: Statistics Versus Good Judgment in Choosing Stent Operators for a Multicenter Clinical Trial
George Howard, Jenifer H. Voeks, James F. Meschia, Virginia J. Howard and Thomas G. Brott

Stroke. 2014;45:3325-3329; originally published online September 11, 2014;
doi: 10.1161/STROKEAHA.114.006807
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2014 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/45/11/3325

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke 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 Stroke is online at:
http://stroke.ahajournals.org/subscriptions/