First-Time Failure Rates of Candidates for Board Certification

An Educational Outcome Measure

Denis M. O'Day, MD; Chun Li, PhD

Background: Few objective standards are available to assess the educational effectiveness of ophthalmology residency programs. As a possible measure, we evaluated the first-time failure (FTF) rate in the examinations of the American Board of Ophthalmology, defined as a first-attempt failure in the written examination or a first-attempt failure in the oral examination after having passed the written examination on the first attempt.

Method: We tracked data on all residents who graduated between June 30, 1999, and December 31, 2003, from commencement of training to certification, including rates of overall FTF, written and oral FTF, and program FTF. Performance was analyzed for several factors, including program size.

Results: The FTF rate was 28% overall and ranged from 0% to 89% across 118 programs (median, 27%). Programs with fewer than 16 graduates per 5 years were significantly more likely to have higher FTF rates than larger programs. Thirty-two programs accounted for 50% of the FTF rate.

Conclusions: The FTF rate is a potentially useful measure. However, the small size of many programs contributes to some imprecision. Therefore, this measure should be used in conjunction with other factors when assessing the educational effectiveness of ophthalmology residency programs. Although the eventual certification rate was high, graduates from a few programs appeared inadequately prepared to take the examinations.

Arch Ophthalmol. 2008;126(4):548-553
attempt or who, having passed the WQE on the first attempt, fail the oral examination on the first attempt. We calculated the FTF rate for the entire cohort of candidates and for individual programs. For the evaluation of program performance, we counted residents who graduated from the program, including transfers into the program. Residents who transferred from a program were excluded from its tally.

**STATISTICAL ANALYSIS**

Suppose program \(i\) \((i = 1, \ldots, 118)\) has a true but unknown failure rate \(\theta_i\). Let \(n_i\) be the number of graduates from program \(i\). Then the number of failures \(x_i\) for program \(i\) can be viewed as a random number from a binomial distribution with \(n_i\) draws and failure probability \(\theta_i\). One estimate of \(\theta_i\) is the raw or naive rate \(\theta_{\text{ni}} = x_i/n_i\). As shown in the “Results” section, the naive rates may have various precision levels, and the rates for smaller programs tend to be underestimated or overestimated by a large margin. Alternatively, in an empirical Bayes method, the 118 \(\theta_{\text{ni}}\)’s are assumed to come from a common distribution, beta \((\alpha, \beta)\), in which the parameters \(\alpha\) and \(\beta\) control the shape of the distribution and may be estimated using the data.\(^4\)\(^2\) For our data, the parameters were estimated to be \(\alpha = 2.84\) and \(\beta = 6.82\). This \(\beta\) distribution reflects the distribution of a program’s possible failure rate and serves as a prior distribution as if it were known before analyzing data. This prior information together with a program’s data are used to derive a new distribution, called a posterior distribution, reflecting updated knowledge about the unknown failure rate for the program. Each program will have its own posterior distribution. Because additional information has been used, the posterior distributions will have smaller variation than the prior distribution and the variation of a large program will be smaller than that of a small program. In this model, the posterior distribution is also a beta distribution.\(^3\)\(^5\)

A program’s FTF rate was estimated by the mean of its posterior distribution (posterior mean). The value lies between the program’s naive FTF rate and the overall FTF rate due to a phenomenon called shrinkage.\(^4\)\(^4\) The magnitude of shrinkage reflects the level of uncertainty in current knowledge about a program’s FTF rate. For large programs, the naive FTF rate and the posterior mean tend to be similar, whereas for small programs, they may differ to a greater extent. Thus, shrinkage effectively pulls back the estimates for small programs that would otherwise have been underestimated or overestimated by the naive estimates. The programs may be ranked by their posterior means. However, when ranking is the primary outcome, the optimal way of ranking a program is based on its expected rank, which is the average of all its potential ranks weighted by their associated probabilities.\(^4\) The expected rank can also show shrinkage.\(^4\) All analyses were performed with the statistical software program R (http://www.r-project.org).

**RESULTS**

From June 30, 1999, to December 31, 2003, 2163 residents graduated from the 118 programs. Eleven did not complete training in the specified time. Of the 2106 taking the examinations, 1491 (70.8%) achieved certification without a failure in either examination (Table 1). Twenty-five passed the WQE but had yet to take the oral examination.

Of those who took 1 or both examinations before the end of 2005, 414 FTFs occurred for the WQE and 176 for the oral examination for a total of 590 FTFs or 28%. Program size ranged from 7 to 42 graduates (Figure 1A). Fifty-nine programs had 15 or fewer graduates and 59 had 16 or more graduates. Eighty-nine programs (75%) had 21 or fewer graduates. The numbers of FTFs varied considerably across programs, from 0 to 17. When the programs are ranked according to the number of FTFs, approximately half came from the bottom 32 programs. The naive FTF rates ranged from 0% to 89% (median, 27%; first and third quartiles, 17% and 41%) (Figure 1B).

Large programs tended to have lower failure rates (Pearson \(\chi^2\) test of independence, \(P = .02\)). We also divided the programs into quarters based on their sizes (Table 2) and calculated the corresponding FTF rates. Thus, for the training programs with 15 or fewer graduates, 36 programs had a naive FTF rate of 27% or worse, whereas, in 23 training programs the rate was less than 27%. In contrast, of training programs with more than 15 graduates, 23 had an FTF rate of 27% or more, while in 36 programs the rate was less than 27% (trend test \(P < .001\)). Because the sizes of programs varied, estimates of their naive FTF rates had different precisions.

In general, variation was higher among small programs than among large programs (Figure 2A). Smaller programs tended to rank near the ends, whereas larger programs did not exhibit this tendency (Figure 2B).

When the posterior mean as an estimate of FTF rate is plotted against program size, the effect of shrinkage is apparent (Figure 2C). Figure 2D shows the plot of the expected rank vs program size. Using this measure, the ranks of small programs were more evenly distributed than those based on naive FTF rates. Compared with Figure 2B, the expected ranks of the small programs tended to be near the center, whereas those of the large programs tended to be similar to the ranks based on the naive FTF rate.

The 4 programs with 0 FTFs were tied in first place with naive FTF rates of 0. However, they had different numbers of graduates (Table 3). The empirical Bayes method could differentiate these programs by considering the different uncertainties that resulted from different program sizes.

The largest program, with 42 graduates and only 1 FTF, was ranked fifth based on the naive FTF rate of 0.024. However, it ranked second based on a posterior mean of 0.074. Its expected rank was 5.7, the highest expected rank. The empirical Bayes method produced a fairer comparison between this and the other programs, making it stand out as the best.

Figure 3 shows the ranks based on naive rates and the expected ranks and their associated 90% confidence.
intervals. Smaller programs tended to have wider confidence intervals than larger programs because of the higher levels of uncertainty. A comparison based on expected ranks showed a tendency for them to be nearer to the center than one based on naive FTF rates.

Ranks based on naive FTF rates and expected ranks agreed well, with a correlation coefficient of 0.99 (Figure 4). Using the expected ranks, small programs were not overly rewarded or punished.

Between June 30, 1999, and December 31, 2003, the closure of 3 programs and the merger of 2 others required trainees to change programs to complete training (involuntary transfer). Some other trainees also voluntarily changed programs (voluntary transfer). Overall, 46 of the 2163 graduates transferred from 29 programs (including the 4 closed or merged ones) into 34 programs. None transferred to a closed or merged program. All or the 14 trainees who involuntarily transferred programs took the examinations. There were 3 FTFs in this group. Five trainees who voluntarily transferred programs took the examinations. There were 3 FTFs in this group. Five trainees who voluntarily transferred programs did not take the examinations. Among the remaining 27 voluntarily transferring trainees who took the examinations, there were 14 FTFs.

Thus, 17 of the 41 transferees who took the WQE or both examinations failed for the first time. This FTF rate of approximately 42% was higher than the overall rate of 28%. The voluntary transferred trainees mainly contributed to this higher failure rate. A test of independence between FTF and voluntary status among the transferred graduates yielded a $P = .096$, presumably owing to the small numbers. When we compared FTF rates between graduates who did not transfer and those who transferred voluntarily, a strong correlation was found between FTF and transfer from a nonclosed or nonmerged program ($P = .006$). Among the 2163 graduates, only 57 had not taken any examination (2.6%) compared with 5 of the 32 voluntary transferees (15.6%). This discrepancy also was statistically significant ($P < .001$).

We investigated the effect of these transferees on a program’s ranking. Among the 34 programs that accepted them, 2 were unaffected because they had not taken any examination, 17 benefited because all trainees passed, and 15 did not benefit (in 12 programs, all transferees failed; in 3 programs, half of the transferees failed).

### Table 2. Numbers of First-Time Failures (FTFs) Grouped by Program Size

<table>
<thead>
<tr>
<th>Quartile</th>
<th>No. of Programs</th>
<th>Range of No. of Graduates</th>
<th>Total No. of Graduates</th>
<th>Total No. of FTFs</th>
<th>FTF Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>30</td>
<td>7-12</td>
<td>296</td>
<td>106</td>
<td>0.358</td>
</tr>
<tr>
<td>Second</td>
<td>29</td>
<td>13-15</td>
<td>413</td>
<td>134</td>
<td>0.324</td>
</tr>
<tr>
<td>Third</td>
<td>30</td>
<td>16-22</td>
<td>567</td>
<td>157</td>
<td>0.277</td>
</tr>
<tr>
<td>Fourth</td>
<td>29</td>
<td>23-42</td>
<td>830</td>
<td>193</td>
<td>0.233</td>
</tr>
</tbody>
</table>

The educational performance of training programs for ophthalmology remains largely unexamined. It is generally conceded that since most candidates eventually achieve certification, measuring rates of certification does not distinguish among programs and is thus not a discriminating educational outcome measure.

Before establishing the Resident Tracking Program in 1998, the ABO was unaware of graduates unless they presented for the examinations. As a consequence, the percentage of graduating residents taking and passing the examinations was unknown. In this cohort, more than 97% applied to take the examinations, confirming the importance of ABO certification from the residents’ perspective. According to a recent Gallup poll, the public believes specialty physicians should demonstrate a credible standard of knowledge and skills. Thus, the degree to which residents are prepared for the ABO examination is an important educational outcome measure for training programs.

The WQE and the oral examination assess different skills (cognitive for the WQE and clinical management for the oral examination). A resident properly prepared should expect to pass both in the first attempt. To be fair to programs, we did not attribute subsequent failures or passes to the program since remediation was beyond its control.
In addition to the quality of the training program, many confounding variables may affect a candidate’s performance. Among these, innate ability and prior education are important. Indeed, it might be argued that the better programs attracted residents who would do well regardless of the teaching environment. Lacking information on resident’s aptitude at the time of entry to programs, we are unable to estimate its effect. However, a study designed to separate the effect of resident training from other factors in cognitive examinations similar to the WQE showed that although achievements before residency are important predictors of the outcome of the examinations, candidates perform better or worse than expected, depending on the quality of the training program.

Validity of ranking is also an issue. Indeed, performance comparison is so important that recently the Journal of Educational and Behavioral Statistics devoted an entire issue to it (volume 29, No. 1, Spring 2004). The simplest forms include league tables in sports and medicine and the magazine rankings of universities. A more sophisticated form is the ranking of hospitals. These rankings tend to have multiple effects, including positive or adverse publicity for the program and a reference for future trainees or customers. Even though these are observational studies, causal interpretation often follows because administrators and legislators are tempted to assign praise or blame on the basis of hard data and true measures. A comparison based on a simple measure of performance, such as the naive FTF rate, can be unfair because it often fails to consider some important factors. For a fairer comparison, many statistical challenges must be overcome, and even then, sophisticated methods cannot guarantee a fair comparison.

These challenges have attracted the attention of many statisticians. We chose the expected rank based on the work of Laird and Louis, who showed its superior-

![Figure 2. Characteristics of the programs for failure rates and the number of trainees. Naive failure rate vs program size (A), rank based on naive failure rate vs program size (B), posterior mean vs program size (C), and expected rank vs program size (D). The vertical lines indicate the separation of the programs into 59 small and 59 large ones.](http://archopht.jamanetwork.com/pdfaccess.ashx?url=/data/journals/ophth/6871/)
Table 3. Effect on Ranking of Using the Empirical Bayes Method to Account for Different Program Sizes\(^a\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Training Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td>No. of graduates in 5 years</td>
<td>29</td>
</tr>
<tr>
<td>No. of first-time failures</td>
<td>0</td>
</tr>
<tr>
<td>Failure rate estimated by posterior mean</td>
<td>0.073</td>
</tr>
<tr>
<td>Rank based on posterior mean</td>
<td>1</td>
</tr>
<tr>
<td>Expected rank</td>
<td>5.9</td>
</tr>
<tr>
<td>Relative position based on expected rank</td>
<td>2</td>
</tr>
</tbody>
</table>

\(^a\)P through S indicate 4 separate training programs.

Both the expected rank and the posterior mean thus reversed the order of the programs derived by the naive FTF rate. These 2 programs also demonstrate the effect of shrinkage as a result of using expected ranking, with movement toward the middle by smaller programs. In fact, no program could have expected to rank near 1 unless its superiority to other programs was strongly established. This was almost impossible unless the program sizes were much bigger.

These data are clustered in the sense that the residents are grouped by programs, and this was considered by estimating program-specific FTF rates. This analysis might seem to be different from clustered data analysis commonly seen in statistical literature because we do not have any other input variables or covariates in addition to the grouping variable that indicates the programs.

With these cautions, our analysis provides a glimpse of the state of ophthalmic education not afforded by examining the overall certification rates. Although the high percentage of graduates who eventually achieved ABO certification is reassuring, 29% were insufficiently prepared at the end of training by reason of deficiencies in both cognitive and clinical management skills as evaluated in the examinations. Twenty-seven percent of the training programs contributed to 50% of these failures.

The wide range in naive FTF rates (0% to 89%) among the programs is striking. This disparity in performance remains large even with expected ranking. As shown in Figure 3, the magnitude of the confidence intervals for the FTF rates provides hints about the difference among programs, especially the top and bottom performers. However, since these programs were not identified before the analysis but through the data, any P value calculated on the same data could potentially inflate the difference. In this study, program size emerges as an important influence not just as a limit on the precision of the ranking but because of its significant association with program quality.
as defined by examination performance. A study of internal medicine graduates also found that candidates from larger programs performed better than those from smaller ones. Even though the effect of size was lessened when other known program characteristics were included, the finding was still significant. However, size should not be construed as causal. Rather, it is a surrogate for additional although, as yet, undefined factors. The group of residents who transferred to another program during training is of special interest. For the 14 whose programs closed, the move was necessary to continue training. The rest performed significantly worse than the entire cohort. Their reasons for transferring are unknown, but if dissatisfaction with the program was the motivation, these data suggest that changing programs may not be an effective response.

Data from this study should be kept in perspective. Because of the small size of many programs, we used aggregated data for a 5-year period. A study of internal medicine programs concluded that this was a defensible strategy to increase precision. However, even with the aggregated data, the total number of candidates from many programs was still small. In such programs, a single high- or low-performing candidate can have a greater effect on its rank position than a similarly performing candidate from a larger program. Although ranking provides a range in performance achievable by most disparate programs, bounded on one side by programs that appear to excel and on the other by those that fall far short, the confidence intervals are wide. Because of this and other uncertainties, assessment of individual programs based on the FTF rate should be approached with caution. We do not believe that the public ranking of training programs is useful. The measures and ranks generated from statistical analyses can provide estimates and insights about the performance differences among programs that can aid improvement, but they should not be relied on to the exclusion of other sources of information. A better way is to use these statistics as part of a global evaluation of a program so that the context and, in particular, the effect of program size are better understood.

Regardless of the precise ranking of programs, the number of candidates who were unable to pass the examinations on the first attempt seems unduly high. Data suggest a degree of unevenness in the quality of training among a few of these ACGME-accredited programs that goes well beyond what might be expected of competent educational institutions. This finding deserves our attention as we consider the quality of ophthalmic education in the United States.

A discussion of the possible association with other variables apart from program size is beyond the scope of this article. However, the clustering of these failures in certain programs but not others provides an opportunity to determine the characteristics that are associated with better outcomes. The goal of performance comparison and ranking is not to reward or punish some programs but to identify the factors that make a program succeed or fail and then to use those factors to engineer improvement.

Submitted for Publication: December 8, 2006; final revision received April 23, 2007; accepted June 25, 2007.
Correspondence: Denis M. O’Day, MD, 8000 Medical Center East, North Tower, Nashville, TN 37232-8808 (denis.m.oday@vanderbilt.edu).
Financial Disclosure: None reported.
Funding/Support: This study was supported in part by a grant from Research to Prevent Blindness.

Previous Presentation: Presented in part at the annual meeting of the American Ophthalmological Society; May 23, 2006; Half Moon Bay, California; and published in the 2006 Transactions of the American Ophthalmological Society; it was subsequently modified during the peer review process of the Archives of Ophthalmology.

REFERENCES