Unrealistic expectations arising from mortality data reported in the cardiothoracic journals

Ani C. Anyanwu, MSc, MD, FRCS^a
Tom Treasure, MD, MS, FRCS^b

Background: This study was undertaken to ascertain whether mortality data in the cardiac surgical literature mirror data reported in national databases.

Methods: This was a review of articles with 50 or more subjects reporting single-center mortality data for coronary artery bypass or aortic or mitral valve replacement published in the three major cardiothoracic surgical journals from 1997 through 2000. Mortality data and trends were examined.

Results: One hundred sixty-nine articles were found (coronary artery bypass, $n = 119$; aortic valve replacement, $n = 34$; mitral valve replacement, $n = 16$). Articles were predominantly case series ($N = 95$), with smaller numbers of comparative retrospective studies ($n = 34$), randomized trials ($n = 29$), and prospective noncomparative studies ($n = 11$). The median mortality figures for these studies were 1.5% (interquartile range, 0.3%-2.6%) for coronary artery bypass, 3.4% (interquartile range, 2.0%-5.3%) for aortic valve replacement, and 4.7% (interquartile range, 2.1%-6.9%) for mitral valve replacement. In contrast, the national registry mortality figures were 2.9%, 4.0%, and 6.0%, respectively, in the United States and 2.6%, 4.5% and 6.3%, respectively, in the United Kingdom. Coronary bypass studies with samples smaller than 100 patients reported lower mortality figures (median 0%) than did those with more than 100 patients (1.8%). Exploration with graphical plots suggested a bias toward reporting and publication of studies with below average mortality.

Conclusions: Particularly for coronary artery bypass, published data tend to underrepresent the risk of death as seen in most centers. Outcomes and magnitudes of effects as reported in these research studies may not be replicable to the same degree in most centers. In particular, extreme caution should be taken in extrapolating results from studies with fewer than 100 patients to larger surgical populations.

From the St George's Hospital^a and Guy's Hospital,^b London, United Kingdom.

Received for publication May 15, 2001; accepted for publication June 13, 2001.

Read at the Eighty-first Annual Meeting of The American Association for Thoracic Surgery, San Diego, Calif, May 6-9, 2001.

Address for reprints: Tom Treasure, MD, MS, FRCS, Cardiothoracic Unit, Guy's Hospital, London SE1 9RT, United Kingdom (E-mail tom.treasure@medix-uk.com).

J Thorac Cardiovasc Surg 2002;123:16-20

Copyright © 2002 by The American Association for Thoracic Surgery.

0022-5223/2002 \$35.00 + 0 12/6/119057

doi:10.1067/mtc.2002.119057

Thirty-day mortality is the most frequently quoted outcome measure for cardiac surgery. Although 30-day mortality may not be an ideal outcome and does have limitations,^{1,2} mortality is a robust and reproducible outcome measure that is undeniably important and is therefore universally quoted as a measure of supposed safety of cardiac surgical techniques. This study tests the hypothesis that published series are biased toward reporting lower mortality figures than are achieved in most centers.

TABLE 1. Median and interquartile range of mortality data from journals

	Literature Median	Interquartile range	Society of Thoracic Surgeons National Database	UK Registry
CABG	1.5%	0.3%-2.6%	2.9%	2.6%
AVR	3.4%	2.0%-5.3%	4.0%	4.5%
MVR	4.7%	2.1%-7.0%	6.0%	6.3%

Mortality data from North American (Society of Thoracic Surgeons National Database) and British registries are shown for comparison.

Materials and Methods

We systematically reviewed all issues of three cardiothoracic journals (*Annals of Thoracic Surgery*, *The Journal of Thoracic and Cardiovascular Surgery*, and *European Journal of Cardiothoracic Surgery*) published in 4 consecutive years from 1997 through 2000. All original clinical articles including any reference to coronary artery bypass grafting (CABG), aortic valve replacement (AVR), or mitral valve replacement (MVR) were selected. Articles with survivals reported formed the study sample. Mortality expressed as a percentage is an unstable statistic as the denominator becomes smaller, so an arbitrary cutoff ($n = 50$) was decided on, and articles reporting on smaller numbers of cases were arbitrarily excluded. Articles were also excluded if a 30-day mortality for isolated CABG, isolated AVR, or isolated MVR could not be reliably ascertained, if the data arose from more than one center, or if the study cohort represented a selected high-risk group (such as patients with endocarditis). Review articles, letters, and meta-analyses were excluded. The remaining articles were assessed by a single observer (A.A.). Data extracted for each article included sample size, the subject of study, mortality data for the specified procedure or procedures, study type (case series, prospective cohort, case-control, controlled trial), country of research, specific end point of study, and year or years in which the operations were performed.

National mortality data for 1998 were obtained from the database of the Society of Thoracic Surgeons for North American data (<http://www.sts.org>) and Society of Cardiothoracic Surgeons of Great Britain and Ireland (<http://www.scts.org>).

Statistical Analysis

A 95% confidence interval (CI) around the mortality for each procedure in each article was derived from the following formula: $95\% \text{ CI} = 1.96 \cdot \sqrt{P(1-P)/n}$, where P is the observed mortality. For zero mortality, the 95% CI was computed according to the method described by Ghosh.³ Comparisons between the median mortality figures for subgroups were performed with the Wilcoxon test.

Results

One hundred sixty-nine articles met inclusion criteria—*Annals of Thoracic Surgery*, 102; *The Journal of Thoracic and Cardiovascular Surgery*, 36; and *European Journal of Cardiothoracic Surgery*, 31. Articles were mainly case

Figure 1. Articles on CABG ranked by mortality. Bars represent 95% CIs.

Figure 2. Articles on CABG ranked by upper 95% CI boundary.

series ($n = 95$, 56%), with the remainder being retrospective comparative studies ($n = 34$, 20%), randomized controlled trials ($n = 29$, 17%) and nonrandomized prospective studies ($n = 11$, 7%). Breakdown by operation was as follows: CABG, 119; AVR, 34; and MVR, 16. The median mortality figures for the operation groups are shown in Table 1. Registry mortality figures are shown for comparison. Because of small numbers for the AVR and MVR groups, further analysis was limited to the 119 articles reporting on CABG.

Mortality figures in the articles on CABG are displayed graphically in Figure 1. Although 91 studies reported a mortality below 3%, the 95% CIs show how imprecise the quoted percentage mortality figures are as estimates from which to generalize. Indeed, ranking based on the upper 95% CI boundary shows that the upper 95% CI boundary exceeded the national mortality data in more than half of the

Figure 3. Funnel plot showing asymmetric relationship between sample size and odds of mortality compared with Society of Thoracic Surgeons National Database mortality. In most studies mortality risk was lower than registry mortality (odds ratio < 1); this was sometimes substantial and in some studies (predominantly smaller studies) patients were as much as 10 times less likely to die than were registry patients.

Figure 4. Upper 95% CI boundaries for articles on CABG reporting no deaths. For articles with fewer than 100 subjects it cannot be certain that true risk of mortality is below 3%.

studies (Figure 2). In only a minority of the articles could mortality be precisely defined; the width of the 95% CI was often wide, exceeding 5% in a quarter of articles. There was no significant association between country of publication, publication type, or primary end point and the reported mortality. Although randomized trials on average reported a lower mortality (1.0%) than did other study types, this association was not statistically significant. The reported mortality was, however, statistically significantly lower for articles with fewer than 100 subjects than for those with a sample size of 100 or more patients (Table 2). This is investigated further in the funnel plot (Figure 3), which shows an uneven

TABLE 2. Associations between article characteristics and reported mortality

	N	Mortality (%)		P value
		Median	Interquartile range	
Country of study				
American	46	1.8	0.4-3.2	.48
European	55	1.1	0.3-2.6	
Study sample size				
<100	26	0	0-1.3	.001
≥100	93	1.8	0.9-3.0	
Primary end points of study				
Included mortality	63	1.2	0.3-2.4	.42
Did not include mortality	56	1.6	0.4-3.0	
Study design				
Prospective	36	1.2	0-2.7	.62
Retrospective	83	1.6	0.4-2.6	

distribution with a paucity of studies reporting mortality figures in excess of national averages. Notably, of the 26 studies with a sample size of 100 or fewer patients, only 2 reported mortality figures in excess of the national average.

Discussion

This analysis suggests that reading the articles in the cardiothoracic literature, particularly for CABG, would lead to a lower expectation of the risk of death than that observed on a national or international basis. We also demonstrated the imprecision surrounding mortality data and showed that although most articles reported operative mortality data below registry averages, only in a minority of cases were these statistically different from registry figures. Several factors may contribute to our finding: (1) Surgeons may report on selected patient groups that are not typical of all patients who undergo the procedure (selection bias). (2) Surgical teams may selectively report only those aspects of their practice in which they have average or below average mortality results (reporting bias). (3) Proponents and enthusiasts of a new procedure or modification are unlikely to publish unless the results look good compared with previous methods. (4) Publications are from academic and expert units, who would be expected to have superior results. (5) Patients enrolled in prospective clinical trials are generally likely to have better outcomes than patients receiving the same treatment outside clinical trials.⁴ (6) Finally, workers with worse than average results may choose not to report their results, or journals may be less likely to publish their data (publication bias).

We believe that selection bias, reporting bias, and publication bias all play prominent roles in the cardiothoracic literature. In the 40 of 119 articles on CABG in which the

authors specifically stated that they reported an entire series without exclusion of any definable high-risk group, the median mortality was 2.5%, which is more in line with registry outcomes. In contrast, in the remaining 79 studies, where patient selection was evident or could not be reliably excluded, the median mortality was 1.0% ($P = .001$). Patient selection is therefore a common factor in many studies reporting lower than expected mortality figures.

Cardiac surgical publications are predominantly based on retrospective case series, which makes them more susceptible to publication bias than are clinical trials.⁵ The paucity of small studies that report above average mortality is a hallmark of publication bias.^{6,7} Thirty-one articles (CABG, 26; AVR, 3; and MVR 2) reported zero mortality figures; 20 of these had fewer than 100 subjects. Of the remaining 11, patient selection was evident in 9. Despite the small sample sizes, none of the articles reviewed presented CIs to alert the reader to imprecision of their estimates. The danger of drawing conclusions about low risk or safety from a zero numerator has been well described and quantified⁸ and is demonstrated in Figure 4, which shows that for some articles reporting zero mortality the true mortality risk could be as high as 7%. We believe that the influence of small sample size would have been even greater had we not excluded studies with sample sizes smaller than 50.

These observations have implications for interpretation and application of cardiothoracic research findings. Operative mortality remains a critical indicator of quality because it provides a measure of safety of a technique. Surgeons looking to adopt new approaches described in the literature will be guided in part by operative mortality rates. Because few studies present results for unselected patients, however, it should not be assumed that results similar to those published in the literature are reproducible. Although we recognize that some superior results represent excellent outcomes in expert units and accept that such studies should represent a standard to which all units should aspire, the overall picture presented in the literature is overly optimistic and not reflected in registry data. Although experimental studies are accepted as the most robust form of clinical evidence, our review of the cardiothoracic literature found many randomized trials that were based on small samples, some with fewer than 50 subjects. These included trials making recommendations on important aspects of cardiac surgical practice, such as myocardial preservation. Demonstration of survival advantage of one method relative to another in routine CABG, however, would require a trial with thousands of patients.⁹ Results from small studies should therefore be interpreted only with full appreciation of their methodologic limitations.

We cannot propose any solutions that will reverse the overrepresentation of studies with below average mortality figures. We believe that this trend will remain and stems

from the professional emphasis placed on publication and from an understandable reluctance to publish one's results if they show above average mortality figures. Although analyses of data from large registries and multiple centers have several advantages and may be more representative, they cannot replace case series, prospective studies, and clinical trials as the preferred modalities for evaluating new techniques. Journal editors and peer reviewers can, however, ensure that if data are potentially misleading this is made clear and preferably accompanied by some measure of precision (CI). Although no solutions have been offered, our observation of disparity between the literature and the registries at least alerts clinicians to the limitations of published data. Caution should particularly be exercised in interpreting results from studies with fewer than 100 subjects, because selection bias and publication bias are more likely in these cases.

References

1. Sergeant P, Meyns B. La critique est aisee mais l'art est difficile. *Lancet*. 1997;350:1114-5.
2. Treasure T. Rational decision-making about paediatric cardiac surgery. *Lancet*. 2000;355:948.
3. Ghosh BK. A comparison of some approximate confidence intervals for the binomial parameter. *J Am Stat Assoc*. 1979;74:894-900.
4. Edwards SJ, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J. Ethical issues in the design and conduct of randomized controlled trials. *Health Technol Assess*. 1998;2:i-132.
5. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337:867-72.
6. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-34.
7. Egger M, Smith GD. Misleading meta-analysis. *BMJ*. 1995;310:752-4.
8. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA*. 1983;249:1743-5.
9. Liu Z, Valencia O, Treasure T, Murday AJ. Cold blood cardioplegia or intermittent cross-clamping in coronary artery bypass grafting? *Ann Thorac Surg*. 1998;66:462-5.

Discussion

Dr T. Bruce Ferguson, Jr (New Orleans, La). I appreciate the invitation to discuss this article and congratulate the authors on a provocative study.

In their examination of the contemporary cardiac surgical literature, the authors have raised an important issue in the assessment of surgical outcomes data. In comparing and benchmarking published outcomes data, it is important to evaluate the study design, the nature of the data set, and the risk adjustment techniques used, if any, in these analyses. As pointed out by the authors, sample size can be a major factor affecting results, even in randomized trials. Study designs may be ranked in terms of validity from prospective randomized trials, to observational data sets, to single institutional series; however, many of these single institutional series are designed to evaluate new technology, or at best one technology against another.

The unique value of the national data sets such as the Society of Thoracic Surgeons National Database and the United Kingdom Registry is that they are large and multicentric and can undergo

sophisticated risk-adjustment analyses not possible with smaller data sets. The Society of Thoracic Surgeons National Database currently consists of 1.65 million patient records collected since 1989. A recent study of more than 1 million patients undergoing isolated CABG at 522 sites in the United States and Canada has documented a 41% decline in relative risk-adjusted operative mortality during the decade from 1990 to 1999. Importantly, this occurred in the face of a relative increase in expected operative mortality of 33% during that decade predicted in a time-trend analysis. Without the multicentric national database structure, this type of analysis is not possible. Moreover, the infrastructure of these national databases allows information exchange and analysis that can directly result in cardiothoracic surgical quality improvement.

I have two questions. First, in your review of the literature were a variety of risk-adjustment techniques found? In particular, what risk adjustment techniques were used in the larger sample size studies shown in your Figure 3? Second, did you find a trend in mortality that moved toward these national database benchmarks when only the larger sample size or risk-adjusted series were analyzed?

Mr Anyanwu. With respect to risk adjustment, we did not take that into consideration, and we reported all articles as they had been reported. Most articles were case series and did not in any way incorporate any form of risk adjustment. Most were evaluating outcomes of a given technique or were randomized trials comparing two techniques, and the mortality data were presented as a raw outcome with no risk adjustment.

With respect to changes in trends in mortality, if we were to consider just the largest studies with thousands of patients the median mortality for those was still below the registry average, at about 1.4%. I think that part of the reason for this is that most of these came from large North American or Australian centers and a European center that are known to have good results. If we were all to publish our results in tens of thousands of patients, however, the median mortality would shift toward the national average.

Dr Brian Buxton (Heidelberg, Australia). I enjoyed your article, and I wonder whether the problem is even greater than you have shown us here. I question the accuracy of some national registries. The reason for raising this issue is that we in Australia had a reported mortality for coronary bypass surgery of about 2% to 2.5% for the national average. Recently we introduced a national death index, by which we linked the surgical procedure with

patients' survival or death in a real fashion. Because most people do not die after leaving the country, I think that the death registry accurately records the true death rate. I was quite concerned to find that the surgeon-reported 30-day mortality figures in our society were only about 50% accurate, that is, quite a few patients were not known by the surgeon to be dead but were found to be dead when checked with the national death index. I wonder about the accuracy of some databases because of their voluntary nature. The only other index I know that is connected with a death registry is the valve registry of Ken Taylor in the United Kingdom, and he found the same thing we did, that only about 50% of the early deaths were recognized by the surgeon. It is worse when you analyze survival data, because most physicians lose track of their patients with time, and I found that surgeons only recognized a third of the patient deaths during a period of 10 years. Would you like to comment?

Mr. Anyanwu. I think if anything that the limitations of the validity of the registry would mean that the registries actually underestimate rather than overestimate mortality. As you said, some deaths go unreported, and so it could be that the true mortality of coronary bypass is not 2.5% but is 3.5%, for example. If that is the case, then the publication bias is even more exaggerated than we demonstrated. Without doubt, however, even if we were to cast aside the issue of the national registry mortality figures and look just at the data in the literature, there is clear evidence of publication bias, as shown by the difference in mortality figures between series of selected patients compared with unselected series, between small samples and big samples, and looking at the funnel plot there is an asymmetric distribution that suggests a wealth of unpublished data that are not being reported.

Dr Jeffrey Gold (Bronx, NY). I compliment you on your excellent presentation. We who practice cardiac surgery in New York State have had the privilege of reporting our results and of having them monitored carefully by the state. This includes the use of the Bureau of Vital Statistics, which tracks every 30-day mortality. These highly monitored outcome numbers track closely with the Society of Thoracic Surgeons National Database statistics on coronary and valvular heart disease. This provides a comparison of a highly audited involuntary system with and that of a voluntary nonaudited system. It would appear that the trends that were reported here are indeed consistent with our observations and are accurate.