Circulation
Cardiovascular Quality and Outcomes

American Heart Association®

Limitations of Ranking Lists Based on Cardiac Surgery Mortality Rates

Sabrina Siregar, Rolf H.H. Groenwold, Evert K. Jansen, Michiel L. Bots, Yolanda van der Graaf and Lex A. van Herwerden

Circ Cardiovasc Qual Outcomes. 2012;5:403-409
doi: 10.1161/CIRCOUTCOMES.111.964460
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 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/5/3/403

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/

Downloaded from http://circoutcomes.ahajournals.org/ at University of Maryland--Baltimore on October 17, 2013

Methods Paper

Limitations of Ranking Lists Based on Cardiac Surgery Mortality Rates

Sabrina Siregar, MD; Rolf H.H. Groenwold, MD, PhD; Evert K. Jansen, MD; Michiel L. Bots, MD, PhD; Yolanda van der Graaf, MD, PhD; Lex A. van Herwerden, MD, PhD

Background—Ranking lists are a common way of reporting performance in cardiac surgery; however, rankings have shown to be imprecise, yet the extent of this imprecision is unknown. We aimed to determine the precision of, and fluctuations in, ranking lists in the comparison of cardiac surgery mortality rates.

Methods and Results—Information on all adult cardiac surgery patients in all 16 cardiothoracic centers in The Netherlands from January 1, 2007, until December 31, 2009, was extracted from the database of the Netherlands Association for Cardio-Thoracic Surgery (n=46883). Ranks were assessed using crude and adjusted mortality rates, using a random effects logistic regression model. Risk adjustment was performed using the logistic EuroSCORE. Statistical precision of ranks was assessed with 95% confidence intervals. Additional analyses were performed for patients with isolated coronary artery bypass grafting. The ranking lists, based on mortality rates in 3 consecutive years, showed considerable reshuffling. When all data were pooled, the mean width of the 95% confidence intervals was 10 ranks using crude and 8 ranks using adjusted mortality rates. The large overlap of the confidence intervals across hospitals indicates that rank statistics were not materially different. Results were similar in the isolated coronary artery bypass grafting subgroup.

Conclusions—Rankings are an imprecise statistical method to report cardiac surgery mortality rates and prone to (random) fluctuation. Hence, reshuffling of ranks can be expected solely due to chance. Therefore, we strongly discourage the use of ranking lists in the comparison of mortality rates. **(*Circ Cardiovasc Qual Outcomes*. 2012;5:403-409.)**

Key Words: cardiac outcomes ■ surgery ■ outcomes research ■ mortality rates

Football leagues, college and university rankings, the Thomson Reuters league tables in business: In all types of branches, teams, institutions, or companies are ranked based on their performance. Ranking lists are convenient in the way they present results: One can see at a glance who is performing well and who is not.

The history of ranking lists in cardiac intervention outcomes goes back to 1987, when the Health Care Financing Administration published Medicare cardiac surgery mortality rates in the United States.[1,2] It was the start of public scrutiny in the field of cardiac surgery, which has continued to exist until today.

Such lists, however, could have major consequences in cardiac surgery and the rest of health care. After all, feedback on institutions, regulatory interventions, marketing strategies, and, of course, the choice of physicians and patients might all be influenced by their results. As an example, previously published ranking lists on cardiac surgery mortality rates in New York State led 20% of the bottom quartile surgeons to relocate or cease practicing within 2 years.[3]

With these potential consequences, ranking lists should be scrupulous; however, rankings have been criticized before for their limitations in the comparison of institutional performance. In 1996, Goldstein and Spiegelhalter showed that ranks are misleading when they are interpreted without taking into account their statistical imprecision (ie, chance variation).[4] The New York State cardiac surgery lists showed massive reshuffling of ranks from year to year. Almost half of the surgeons had moved to the other half of the ranking list in 1 year, suggesting a substantial impact of chance variation[5]; however, it is unknown to what extent ranking lists differences can be attributed to real differences or are merely reflecting random variation.

Since cardiac intervention outcomes are increasingly being evaluated using peer-comparison, it is crucial to know whether ranking lists are the suitable format to do so. Although the statistical limitations have been discussed before,[4] ranking lists of cardiac intervention outcomes have never been evaluated before using patient data. Since 2007, the Netherlands Association for Cardio-Thoracic Surgery has collected data on all adult cardiac surgery. Using this clinical database, our aim was to determine the precision of, and fluctuations in, ranking lists in the comparison of cardiac surgery mortality rates across centers.

Received December 5, 2011; accepted April 10, 2012.

From the Department of Cardio-Thoracic Surgery (S.S., L.A.V.); Julius Center for Health Sciences and Primary Care (R.H.H.G., M.L.B., Y.V.), University Medical Center Utrecht, Utrecht, the Netherlands; Department of Cardio-Thoracic Surgery, Institute for Cardiovascular Research, VU University Medical Center, Amsterdam, the Netherlands (E.K.J.).

Correspondence to Sabrina Siregar, MD, University Medical Center Utrecht, Heidelberglaan 100, E03.511, PO Box 85500, 3508GA Utrecht, The Netherlands. E-mail s.siregar@umcutrecht.nl

© 2012 American Heart Association, Inc.

Circ Cardiovasc Qual Outcomes **is available at http://circoutcomes.ahajournals.org** **DOI: 10.1161/CIRCOUTCOMES.111.964460**

Table 1. Characteristics of Data Set

	2007 N=15195	2008 N=15776	2009 N=15912	All Years N=46883
CABG	71.8% (7.7)	69.8% (8.3)	68.4% (8.4)	69.8% (7.8)
Isolated CABG	54.6% (8.8)	54.2% (8.7)	52.8% (8.4)	53.8% (8.3)
Valvular surgery	38.5% (7.8)	38.2% (8.5)	39.1% (7.0)	38.6% (7.5)
Valve and CABG	14.8% (2.9)	13.8% (2.6)	13.9% (2.2)	14.1% (2.2)
Logistic EuroSCORE	7.1% (1.1)	7.0% (0.9)	7.1% (1.0)	7.1% (0.9)
Mortality	3.1% (0.8)	3.3% (0.8)	2.7% (0.8)	3.1% (0.6)
Center volume	446–1953	523–1933	583–1969	1599–5730
	621/843/1133	721/837/1165	745/857/1136	2176/2388/3434

Analyzed on hospital level: Values indicate the means of 16 hospitals, with standard deviations between brackets. Volumes are reported as ranges and quartiles.

Methods

Data

Information was extracted from the database of the Netherlands Association for Cardio-Thoracic Surgery. All records of adult patients undergoing cardiac surgery in all 16 cardiothoracic centers in The Netherlands from January 1, 2007, until December 31, 2009, were used, which comprised 46 883 surgical procedures. The dataset consisted of demographic characteristics, details on the intervention, in-hospital mortality, and risk factors for mortality after cardiac surgery, notably EuroSCORE variables.[6]

Within-Hospital and Between-Hospital Variability

When a variable is compared across hospitals, 2 sources of variability must be distinguished: variability due to chance (ie, within-hospital variability) and variability due to systematic differences between hospitals (ie, between-hospital variability). To study these, distribution plots of the variable can be drawn from the collected data for each center. These can then be used to calculate the mean for each center and its corresponding 95% confidence interval. Accordingly, one can decide whether the differences seen across centers can mainly be attributed to within-hospital variability or to between-hospital variability. For example, wide and overlapping distribution plots or confidence intervals indicate large within-hospital variability and small between-hospital variability.

Confidence Intervals Around Ranks

Contrary to a variable such as age, distribution plots and confidence intervals cannot be constructed from the data in such a straightforward way for ranking statistics. In order to do so, a simulation technique called bootstrapping had to be used, which is a flexible way of evaluating the random variation in empirical data.[7] This means that within each center, samples as large as the original sample were drawn from the database with replacement. Resampling was performed 1000 times, thus yielding 1000 simulated databases. A ranking list was constructed in each of these new databases, resulting in 1000 simulated ranks for each center. The distributions of the simulated ranks were then used to calculate the mean rank and 95% confidence interval (ie, interval between 2.5% and 97.5% quintiles) for each center.

Analysis

Ranking lists for each year (2007, 2008, and 2009) were constructed based on crude mortality rates, as well as risk-adjusted mortality rates using the logistic EuroSCORE. For the latter, a random effects logistic regression model was used, with mortality as the dependent variable, the logistic EuroSCORE as independent variable, and hospital as grouping factor. This random effects model accounts for within-hospital variability and between-hospital variability.[8–10] Hospitals were ranked according to their random intercepts, which reflects the between-hospital variation. We updated the logistic EuroSCORE model in our data by including it in the regression model as a dependent variable (equivalent to an adjustment of the original intercept).[11]

To assess the precision of ranks, all data of 2007, 2008, and 2009 were combined, and bootstrapping was applied. This resulted in mean ranks and accompanying 95% confidence intervals for each center.

To investigate the possible effects of the heterogeneity of procedures on the precision of ranks, all analyses were repeated in a subgroup of only isolated coronary artery bypass grafting (CABG) procedures.

Results

Table 1 shows the types of cardiac surgery included in the database, the mean logistic EuroSCORE, and the mortality rates for all years separately and combined. Approximately half of the interventions were isolated CABG, and more than a third of all procedures involved valvular surgery. The mean mortality rate over 3 years was 3%, and the mean EuroSCORE was 7%. Hospital volumes ranged from approximately 500 to 2000 patients per year and roughly 1600 to 5700 for the 3 years combined. When hospital volume was included in the benchmarking model, no significant effect was found (when categorized into 5 volume classes, all probability values were above 0.15), which suggests that volume had no effect on the risk of mortality in our data.

Figure 1 shows the ranking lists based on risk-adjusted mortality rates for the separate years 2007, 2008, and 2009. Reshuffling of ranks is observed across the years using both methods: not a single hospital maintains its rank throughout the 3 years.

The distributions of the simulated ranks are presented in Figure 2. There is large overlap in the distributions of ranks both when crude and adjusted mortality rates are used. A few narrow and peaked curves at the high and low end of the ranking list can be distinguished in the figure; however, most hospitals contribute to the agglomeration of curves in the wide middle segment of the plot. This illustrates that the highest and lowest ranked hospitals are consistently ranked in high and low positions, respectively, despite random variability (due to chance); however, most hospitals are in the middle part of the ranking lists, where the flat and wide distribution curves indicate that the hospital ranks are likely to fluctuate due to chance.

The distribution plots of ranks can be translated into 95% confidence intervals, as seen in Figure 3. Wide intervals are

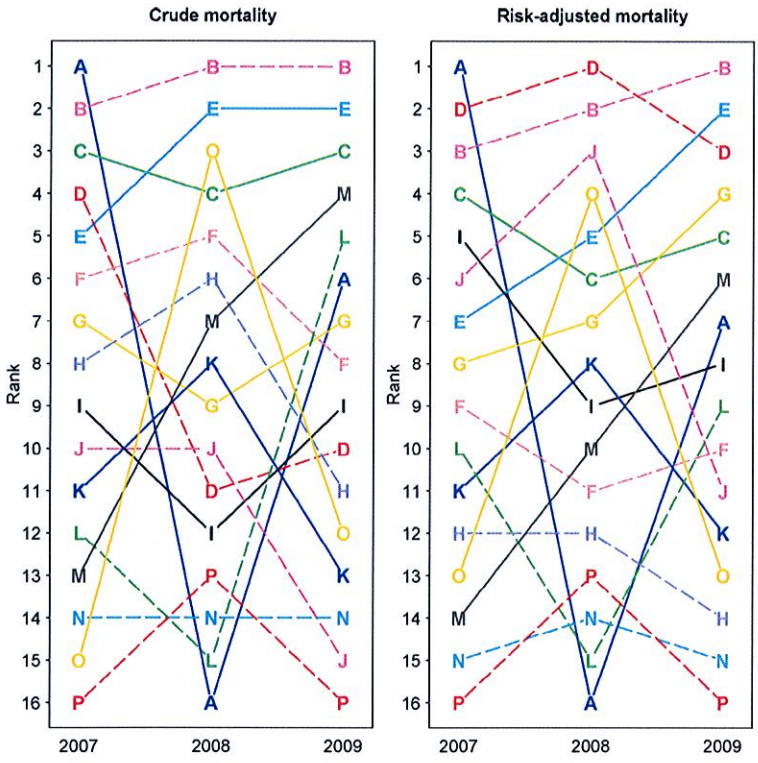

Figure 1. Ranking lists based on crude and risk-adjusted mortality rates of all 16 cardiothoracic surgery centers in The Netherlands for the years 2007, 2008, and 2009, separately. Risk-adjustment was performed using a random-effects model. Reshuffling of the ranks is seen across the years in both panels.

seen with much overlap across hospitals. For the ranking list constructed using crude mortality rates, the average width of all confidence intervals was 10 ranks and for the lists using risk-adjusted mortality, 8 ranks. This indicates that the ranks are imprecise. The 2 highest ranked hospitals have ranks that significantly differ from the 2 lowest ranked, because the confidence intervals do not overlap. As with the distribution plots in Figure 2, this means that the hospitals in the top of the

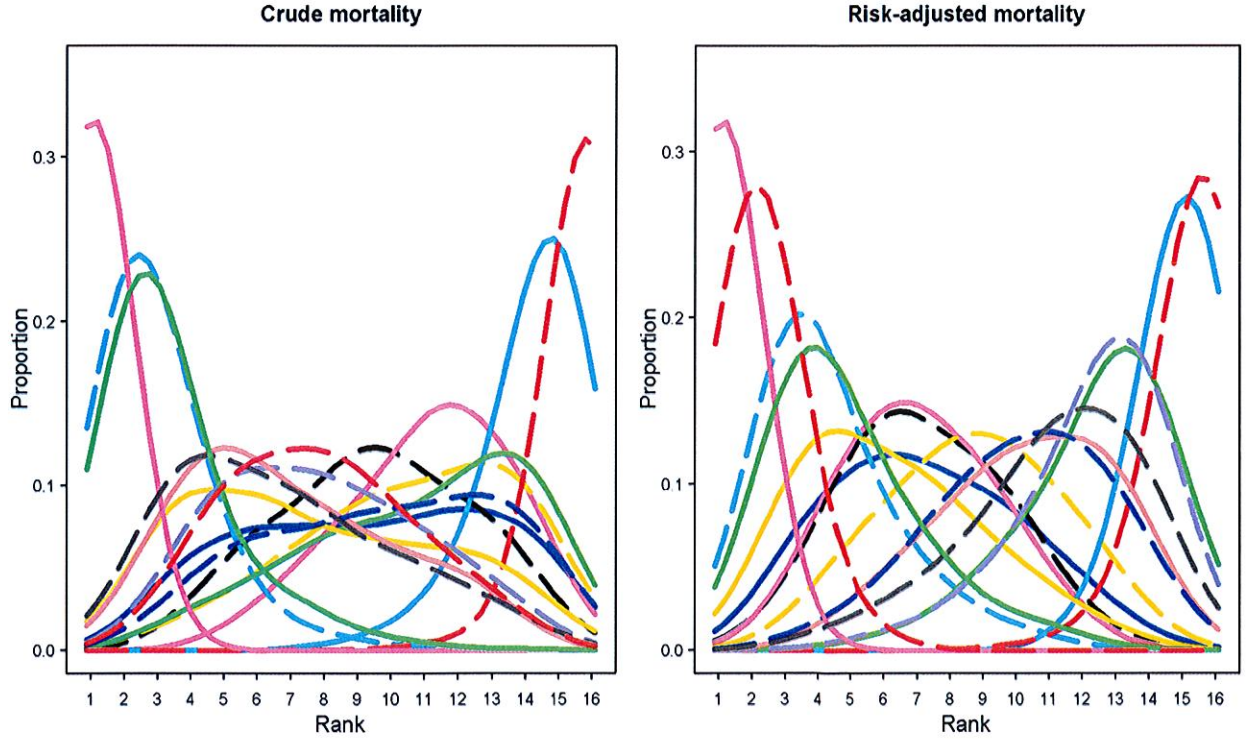

Figure 2. The distributions of the ranks of all 16 cardiothoracic surgery centers in The Netherlands using pooled data from 2007 to 2009. Each curve represents the distribution of the simulated ranks in 1 center. Ranks are assessed with crude mortality rates and adjusted mortality rates (using a random effects logistic regression model). Much overlap in the distribution of the ranks is seen, indicating that most ranks do not significantly differ.

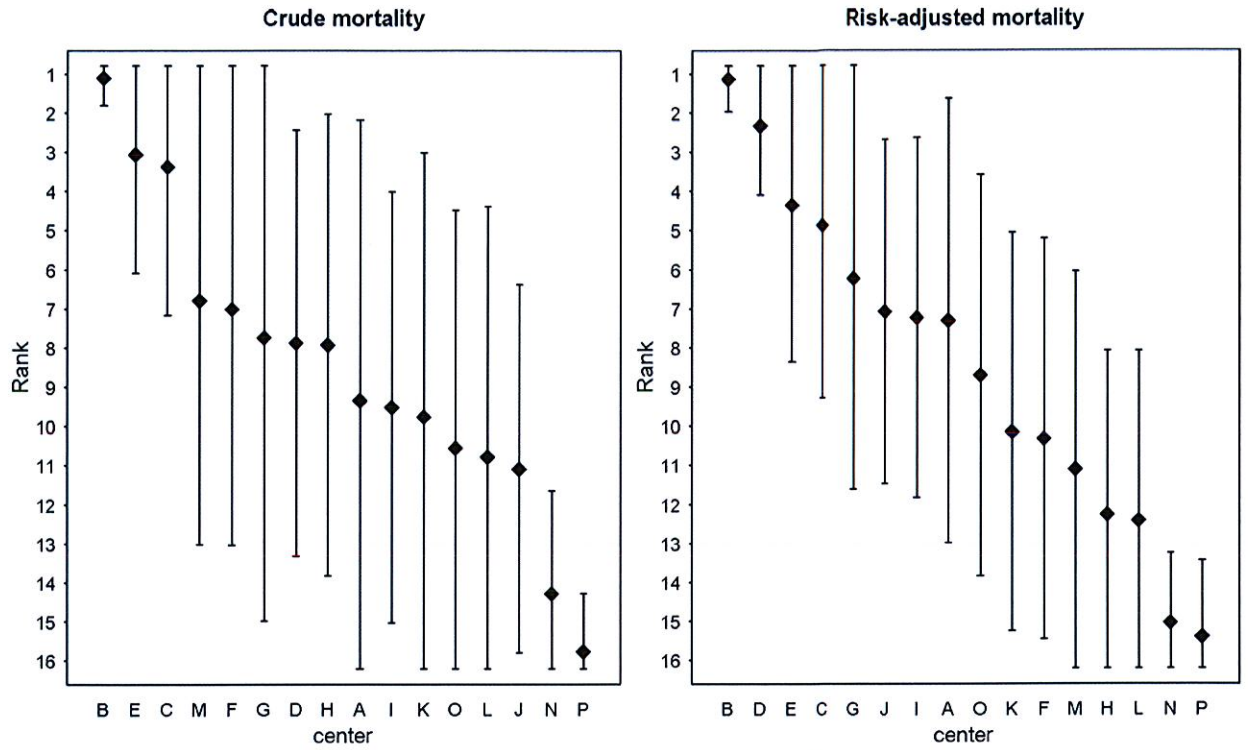

Figure 3. Ranks of all 16 cardiothoracic surgery centers in The Netherlands and their 95% confidence intervals using pooled data from 2007 to 2009. Each point represents 1 center; bars indicate 95% confidence intervals. Ranks are assessed with crude mortality rates and adjusted mortality rates (using a random effects logistic regression model). Much overlap in the confidence intervals of the ranks is seen, indicating that most ranks do not significantly differ.

list are not likely to end up in the bottom of the list and the other way around, merely on account of chance; however, all other ways of reshuffling of ranks is very likely to happen, due to chance variability, because of the strong overlap of confidence intervals.

We identified 25 095 patients with isolated CABG performed from 2007 until 2009. Subgroup analysis with only patients with isolated CABG showed similar results. Large confidence intervals of the ranks were seen when crude mortality and adjusted mortality were used: The mean width of the confidence intervals was 11 ranks, as shown in Figure 4.

Discussion

Principle Findings

We used data on cardiac surgery in all 16 centers in The Netherlands to investigate the precision of ranking lists of cardiac surgery mortality rates. This study showed that ranking statistics were very imprecise. Ranks were likely to fluctuate merely due to chance and were thus instable. The results held true for both crude and risk-adjusted mortality rates.

Statistical Imprecision and Relativity of Ranks

When mortality rates are considered, a distinction must be made between variability caused by systematic differences (between-hospital variability) in the mortality rates and that caused by chance (within-hospital variability). If this chance variability is not taken into account, differences between hospitals are exaggerated and do not reflect the true between-hospital variability.

In addition to this within-hospital variability, rank statistics have another source of variability due to chance: correlation between ranks. Therefore, the confidence intervals of ranks are even wider than that of mortality rates, which are not correlated. The problem is best illustrated by the following example: When a hospital moves from rank 6 to rank 1, all hospitals ranked from 1 to 5 will go down 1 rank even without any changes in the underlying mortality rates. In other words, in a ranking list, a hospital can move in rank without any change in the underlying mortality rate but only because another hospital changed. This also means that the mortality rate of a center is always directly compared with other hospitals (relative scale) and cannot be interpreted on its own (absolute scale). Even when a center has a significantly higher or lower rank than other centers, this merely indicates a relative performance. High and low ranks do not necessarily imply absolute high or low performance. Moreover, it has been shown that this form of direct comparison of hospitals is only valid when case-mix between hospitals is comparable and should otherwise not be performed.[12]

The width of the confidence intervals represents the extent of chance variation that should be taken into account. The large confidence intervals of ranks thus indicate a large amount of random variation, which is likely to cause reshuffling of ranks merely by chance. In other words, the statistical imprecision of ranks causes the ranks to reflect random variation instead of systematic differences in mortality rates.

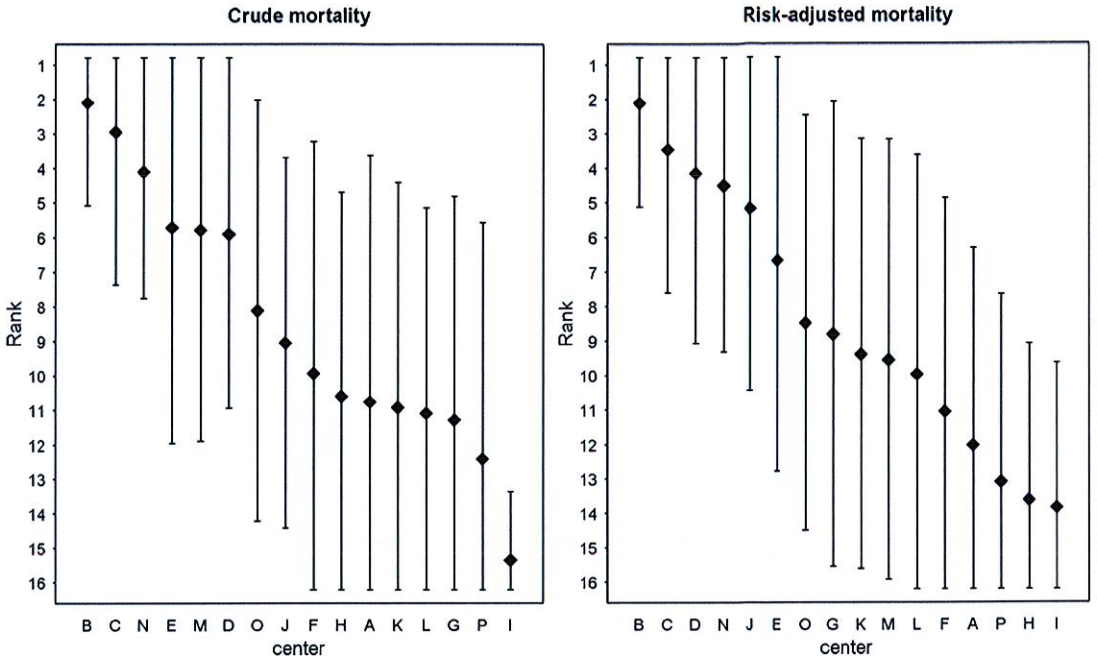

Figure 4. Ranks of all 16 cardiothoracic surgery centers in The Netherlands and their 95% confidence intervals using pooled data from 2007 to 2009, based on isolated CABG procedures only. Each point represents 1 center; bars indicate 95% confidence intervals. Ranks are assessed with crude mortality rates and adjusted mortality rates (using a random effects logistic regression model). As with all procedures combined, much overlap in the confidence intervals of the ranks is seen, indicating that most ranks do not significantly differ.

Without notion of the imprecision of the estimation, one would be unaware of the fact that most values actually do not differ significantly. By definition, a ranking list requires centers to be ranked even when the differences are negligible. Hence, simple ranking lists ignore both the uncertainty around the estimates as well as the magnitude of the differences.

Previous studies on this topic concluded similarly. Ranstam showed that even small amounts of missing data can considerably increase the margin of error around ranks, and Feudtner found wide confidence intervals around ranks based on mortality rates as well.[13,14]

When only 1 certain type of procedure was analyzed (in this case, isolated CABG procedures), results were similar. The average confidence interval was even slightly larger compared with those resulting from all procedures because of the smaller sample sizes. This indicates that the fluctuations and imprecision of ranks cannot be accounted to the heterogeneity of the population.

Consequences for the Use of Ranking Lists

The extent of imprecision and fluctuations of ranking lists depend on the sample sizes and the differences in the underlying mortality rates between hospitals. Referring to the first, nearly 47 000 procedures in 16 centers over a period of 3 years were included in our study. In reality, even larger sample sizes are hard to realize, and more stable ranking lists will be difficult to accomplish for that reason alone. In addition, the variation of hospital volumes is not likely to have affected our results, considering the fact that hospital volume had no significant effect when it was included in the benchmarking model.

Referring to the second, larger differences between the hospital mortality rates will likely result in less overlap of distributions and confidence intervals. For example, in the highest and lowest ranked hospitals there was a large difference in the underlying mortality rate, which resulted in fairly stable ranks. One could hypothesize that in a population with greatly diverging mortality rates between hospitals, ranking lists could be more stable than our results might suggest. The relation between within-hospital and between-hospital variance is described as *rankability* by Van Dishoeck.[15] Rankability is large when the differences between hospitals dominate the within-hospital variance. Yet, even in that case, our general conclusions would hold: (1) The interpretation of ranking lists always requires knowledge about variability due to chance, because it enables to ascertain systematic differences rather than random variation; and (2) chance variability is larger in ranking statistics than in mortality rates, because ranks represent a relative scale and are correlated to each other. Considering the fact that simple ranking lists are never reported with confidence intervals or any other unit to describe precision, we strongly discourage their use in the comparison of cardiac surgery mortality rates. The importance of reporting the margin of error around rank estimates is emphasized by other authors as well.[13,14,16–19] Misinterpretation or plain negligence of the uncertainty surrounding ranks or any other measure will lead to flawed conclusions. Considering the unmerited consequences this might have for some centers, this must be avoided at all times.

Alternatives to Ranking Lists

The limitations of ranking lists should not be an impediment to outcomes evaluation and provider profiling. Whether

outcomes are publicly reported or compared in peer-confidentiality, data collection and feedback seem to be associated with improved outcomes and should therefore be accomplished.[20,21] Fortunately, other ways have been opted to avoid the issues inherent to ranking yet still report differences in mortality rates. Lingsma and Steyerberg propose to use expected ranks.[22] These are rank statistics based on the probability that a hospital performs worse than any other hospital in the ranking list. The expected ranks incorporate the magnitude of difference and thus allow subtle differences between hospitals (eg, rank 4, 5, and 6 versus expected rank 4.1, 4.2, and 5.9). The advantage of this type of ranking is that the probabilities of hospitals performing better than any other (ie, the expected ranks) are not as strongly correlated as usual ranks. When the performance of 1 hospital changes, the expected ranks of other hospitals do not necessarily have to shift as well. This makes the expected rank not so much a rank but another derived measure to compare hospitals. The disadvantage of expected ranks is that statistical imprecision cannot be read from the ranks, nor is it shown as a confidence interval. Again, this makes it difficult to ascertain systematic differences, rather than random variation. Similar to usual ranks, expected ranks enable evaluation of outcomes in a relative way but not in an absolute way.

A more common approach to avoid ranking lists is to compare each hospital against 1 value. This method is based on the identification of statistical outliers. For example, the STS national database uses a 3-tiered rating system for its composite quality scores. Usually, identification of outliers is achieved by assessing confidence intervals of mortality rates and determining their overlap with the overall average mortality rate. When no overlap exists, the mortality rate is significantly different from the overall rate, and the concerning hospital is considered to be an outlier.[12] Other, more statistically advanced techniques include Bayesian analysis to investigate statistical difference with an overall value. When the main goal of evaluating mortality rates is quality control and improvement, these types of methods might be more suitable than ranking lists.

Possible Limitations

Because the goal of this study was to investigate the stability of ranking lists and not to find the optimal approach to compare hospitals, many other issues in the comparison of hospital-specific mortality rates were not discussed. This complex subject is extensively discussed in many other papers.[20] The major concerns are in the area of risk-adjustment models, differences in treated patients (case-mix), and unmeasured risk factors.

The importance of risk adjustment was apparent with the major reshuffling of ranks when crude mortality rates were adjusted for risk. Because the logistic EuroSCORE model is known to have a poor calibration, we recalibrated the model in our data and achieved adequate model performance[23]; however, it can be debated whether the EuroSCORE model is the best method for risk adjustment and whether unmeasured risk factors have caused differences in mortality rates as well. Although much discussion continuous on this topic, there is no reason to assume that another risk adjustment model would lead to different conclusions concerning the large fluctuations in ranks. Both unadjusted as adjusted mortality rates yielded the same results in this matter.

Conclusion

In conclusion, rankings are an imprecise statistical method to report cardiac surgery mortality rates. The 95% confidence intervals of most ranks in the ranking list strongly overlap. As a consequence of this, reshuffling of ranks can be expected solely due to chance, and this was indeed observed over a period of 3 years. Therefore, we strongly discourage the use of ranking lists for the purpose of comparison of risk-adjusted cardiac surgery mortality rates.

Sources of Funding

The Department of Cardio-Thoracic Surgery UMC Utrecht has received financial support from the Netherlands Association of Cardio-Thoracic Surgery to cover part of the first author's salary.

Disclosures

None.

References

1. Adult cardiac surgery in New York State 2005–2007. Albany, New York: New York State Department of Health; 2010.
2. Hospital Guide 2010. What makes a good hospital? *Dr Foster Intelligence*. 2010.
3. Jha AK, Epstein AM. The predictive accuracy of the New York State coronary artery bypass surgery report-card system. *Health Aff (Millwood)*. 2006;25:844–855.
4. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc Ser A Stat Soc*.1996;159:385–443.
5. Green J, Wintfeld N. Report cards on cardiac surgeons. Assessing New York State's approach. *N Engl J Med*. 1995;332:1229–1232.
6. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg*. 1999;16:9–13.
7. Metropolis N, Ulam S. The Monte Carlo Method. *J Am Stat Assoc*.1949; 44:335–341.
8. Lingsma HF, Steyerberg EW, Eijkemans MJ, Dippel DW, Scholte Op Reimer WJ, Van Houwelingen HC. Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. *QJM*. 2010;103:99–108.
9. Thomas N, Longford NT, Rolph JE. *A Statistical Framework for Severity Adjustment of Hospital Mortality Rates*. Santa Monica, Ca: Rand; 1992.
10. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med*. 1994;13:889–903.
11. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61:76–86.
12. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation*. 2008;117:1955–1963.
13. Ranstam J, Wagner P, Robertsson O, Lidgren L. Health-care quality registers: outcome-orientated ranking of hospitals is unreliable. *J Bone Joint Surg Br*. 2008;90:1558–1561.
14. Feudtner C, Berry JG, Parry G, Hain P, Morse RB, Slonim AD, Shah SS, Hall M. Statistical uncertainty of mortality rates and rankings for children's hospitals. *Pediatrics*. 2011;128:e966–e972.
15. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf*. 2011;20:869–874.

16. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ*. 1998; 316:1701–1704.
17. Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group. *BMJ*. 1998;316:1931–1935.
18. van Dishoeck AM, Looman CW, van der Wilden-van Lier EC, Mackenbach JP, Steyerberg EW. Displaying random variation in comparing hospital performance. *BMJ Qual Saf*. 2011;20:651–657.
19. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? *Med Care*. 2005;43: 1177–1184.
20. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, Dreyer PI. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155–2168.
21. Hannan EL, Sarrazin MS, Doran DR, Rosenthal GE. Provider profiling and quality improvement efforts in coronary artery bypass graft surgery: the effect on short-term mortality among Medicare beneficiaries. *Med Care*. 2003;41:1164–1172.
22. Lingsma HF, Eijkemans MJ, Steyerberg EW. Incorporating natural variation into IVF clinic league tables: the expected rank. *BMC Med Res Methodol*. 2009;9:53.
23. Siregar S, Groenwold RH, de HF, Bots ML, van der GY, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg*. 2012;41:746–754.