ASWAN HEART CENTRE
SCIENCE & PRACTICE SERIES
A Qatar Foundation Academic Journal

OPEN ACCESS

Editorial

Statistics re-analysed: From complexity to simplicity and utility

Johanna JM Takkenberg[1,*], Ismail El-Hamamsy[2] and Magdi H Yacoub[3]

[1]University Medical Center Rotterdam, 's Gravendijkwal 230, 3015 CE Rotterdam, Netherlands
[2]Department of Cardiac Surgery, Montreal Heart Institute, Montreal, Canada
[3]Qatar Cardiovascular Research Center, Doha, Qatar
*Email: j.j.m.takkenberg@erasmusmc.nl

"Everything should be made as simple as possible, but not one bit simpler" (Einstein)

There is no shadow of doubt that statistics are an essential part of science and clinical practice, and therefore should be readily available to all concerned, in a comprehensible manner. In recent years, statistics have tended to deviate from this intention, with their perception of being complex and user-unfriendly. This misconception requires an urgent re-evaluation of statistics, with the aim of restoring their original image of simplicity, elegance and application.

In this issue of the Journal, the two excellent articles by Marc de Leval & Ben Bridgewater and Sir Bruce Keogh serve the purposes of clarifying several aspects relating to the philosophy and applications of statistics. We strongly endorse their message, and attempt in this Editorial to present a simplified road map of the applications of statistics including the origins and utility of commonly used methodologies.

HISTORICAL PERSPECTIVE

In 1978 entomologist David W Roubik reported in *Science* on the association between the number of Africanized honey bees and the number of stingless bees on flowering *Melochia Villosa* [1]. His paper contained a scatter plot that not only displayed the actual observations that he made, but also a quadratic polynomial regression line to show the general trend in the scatter plot. This regression line, obviously a "one bit simpler" attempt to model reality, became famous when Robert M Hazen commented on the figure and actually proposed an alternative interpretation of the data (Fig. 1) [2]. This classical example illustrates the challenge that we are still facing in the 21st century when we use statistical methods to explain the complexity of our observations.

To complicate matters: In the past four decades, analog to many other scientific fields, the field of biostatistics has made a giant leap, owing a lot to the ongoing tremendous computational power increase in the computer era. Ironically, "cross-pollination" with medical sciences is lagging, causing medical scientists to often report great medical breakthroughs with outdated and/or incorrectly employed statistical methods.

PITFALLS

Although published research goes through a rigorous process of peer-review to ensure its quality and adequacy, several pitfalls remain with regards to statistical methods which could affect data interpretation.

External validity of a study

All controlled studies require strict inclusion and exclusion criteria in order to homogenize patients' characteristics to a degree that would reduce the likelihood of unaccounted factors influencing outcome in the different study groups. Thus, although prospective randomized controlled trials currently represent the most powerful method for answering clinical questions, most of these trials have long lists of inclusion and exclusion criteria which are aimed at homogenizing the study groups and reducing the likelihood of confounding factors. Although this has the benefit of strengthening the

DOI: http://dx.doi.org/10.5339/ahcsps.2011.14
Published: 30 December 2011

QSCIENCE.com
Member of Qatar Foundation

Cite this article as: Takkenberg JJM, El-Hamamsy I, Yacoub MH. Statistics re-analysed: From complexity to simplicity and utility, *Aswan Heart Centre Science & Practice Series* 2011:14
doi: http://dx.doi.org/10.5339/ahcsps.2011.14

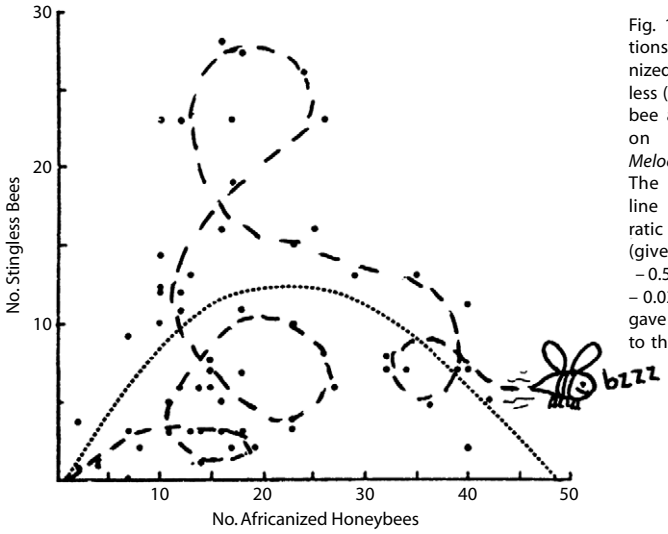

Fig. 1. The relations of Africanized and stingless (meliponine) bee abundances on flowering *Melochia villosa*. The dashed line is a quadratic polynomial (given by $y = -0.515 + 1.08x - 0.023x^2$) which gave the best fit to the points (7).

Figure. 1 Robert M Hazen's observations on curve fitting. Reproduced with permission from Ref. [2].

similarities between groups, the list of exclusion criteria is often quite wide and includes patients with more severe forms of disease or ones with comorbidities. Thus, as these criteria become more restrictive or "exclusive", generalizing the results to daily practice (external validity) becomes more difficult. An example is excluding patients with reduced ejection fraction from studies on valvular surgery. All too often, results will be extrapolated to all categories of patients with varying ejection fractions, yet strictly speaking, if the real-life patient does not meet the inclusion/exclusion criteria of the study, it is less accurate to apply the results.

Publication bias

Rarely do different studies on the same issue unanimously provide similar results. Most often, there are diverging results, varying from the positive findings (e.g. a significant difference between groups or desired effect of a certain drug or procedure) to absence (or even negative) effect. It has often been suggested that positive findings have a higher likelihood of publication in higher impact factor journals than negative studies, even if both studies are have similar methodological merit. It is always important to be reminded that "good" journals do not automatically make "good" papers and the opposite is true. Thus, it is important to always critically scan the entire literature for all available data, both positive and negative, in higher and lesser impact factor journals, before formulating an opinion over a specific question.

In addition, a recent comment in the *New England Journal of Medicine* examining systematic reviews completed by the Cochrane Heart Group found that studies named with an acronym had a higher likelihood of being published in high impact factor journals (namely *Circulation*, *The Lancet* and the *New England Journal of Medicine*) [3]. Furthermore, these studies were cited twice as many times as non-acronym-based studies despite the fact that they were not more likely to report positive results. The role of acronym-based studies as a mnemonic tool is likened to cognitive phenomenon known as automatic attitude activation. A word of caution should be mentioned however since these acronym-named studies are four times as likely to be funded by pharmaceutical industries, and importantly, eight times as likely to be authored by industry employees.

Relative and absolute risk

Study findings are often presented in a manner aimed at capturing the reader's attention. One of the ways of reporting differences between two study groups is to express them in relative terms, i.e. relative risk. This can sometimes lead to striking results. For instance, a reduction in mortality from 1 in 1000 to 0.5 in 1000 represents a 50% relative risk reduction, a significant difference by any standards. Nevertheless, in absolute terms, this means that the number needed to save 1 life is 2000 cases, which may appear less dramatic. Industry-driven data will often be presented in relative terms in order to promote specific drugs or treatment options. Yet, in an era of increasing health-care

expenditures and limited resources, it is always important to reassess the impact of such treatments in absolute terms looking at the "number needed to treat", which may yield marginal benefits from changing practices, and lead to better resource allocation.

ASKING THE RIGHT QUESTION

> "*Not everything that counts can be counted, and not everything that can be counted counts*" (Einstein)

The first and arguably most important part of using statistics is to ask the right question. The renowned statistician, Sir David Cox, stated in one of his seminal lectures at the NHLI: "In the long run it is certainly better to find the wrong answer to the right question, than the right answer to the wrong question". It follows therefore that time spent by the researcher in clarifying the exact question being asked and the reasons why, is time well spent.

In a recent Editorial in *Circulation* entitled "Right Answer, Wrong Question" [4], the author comments on a publication in the same issue of the journal by the CONFIRM investigators who reported a poor correlation between "symptoms" and CT angiography finding of coronary artery disease [5]. In his excellent editorial George Diamond states "At the risk of seeming to construct a straw man argument, ask yourself 'Do symptoms matter?' Is a patient with chest pain more likely to have coronary disease than an asymptomatic patient?" Of course symptoms do matter, and are almost certainly more sensitive in detecting disease, than the most sophisticated test or group of test as pointed out by Carabello in the article published in this issue of the Journal [11]. Thus it is vital to ask the right question when designing a study or interpreting the results of published work.

From the pitfalls described above we can learn a lot about asking the right question and about selecting the correct statistical tool to answer that question. Once we have clearly defined our goal and we need to carefully consider our dataset/patient population and based on the type of data determine what type of statistical tools are most appropriate to answer the question.

TOOLBOX

What does the 21st century statistical toolbox for cardiovascular professionals offer? Randomized controlled trials are the gold standard in clinical research, but these are extremely rare in surgical practice [6]. In the absence of randomized controlled trials, we usually deal with (preferably prospective) patient cohort studies. Simple descriptive statistics remain of crucial importance to paint the picture of our patient population of interest: sample size, patient selection, patient history and characteristics, procedural details, and early morbidity and mortality. Once this picture is clear, we can appreciate the context of the specific patient population and apply more advanced statistics.

Table 1 provides a road map for statistical tests that we can use to answer our research questions. It shows that besides a clear-cut definition of the goal of our study, we need to take into account the type of data that we want to analyze. By cross-tabulating our study goal and the type of data, as detailed in Table 1, we can adequately select the statistical test that is appropriate to answer our question. For example, if we want to study whether patient age influences operative mortality, our study goal is to predict operative mortality using patient age and the type of data that we are using is binomial (two possible outcomes, namely dead or alive). From Table 1 we then learn that the appropriate statistical test is simple logistic regression.

There are several easily available statistical methods to study patient outcome, for example early outcome assessment through logistic regression analysis, and late outcome analysis by means of Cox regression analysis. In the field of cardiothoracic surgery, events that take place early after surgery, usually the first postoperative month, are modeled using logistic regression analysis. Logistic regression analysis has found widespread use in the cardiovascular surgery field through the STS score (www.sts.org) and Euroscore (www.euroscore.org), both risk stratification models that are used worldwide for individual and group surgical risk prediction, and for bench marking. Cox regression analysis is a type of logistic regression analysis that models the time to an event instead of simply an event (as in logistic regression analysis), and is usually employed to study factors associated with time-related event occurrence like death or reoperation over a period of years. It can provide the clinician with important insight into determinants of outcome over a longer time period. For example Table 2 displays the univariable and multivariable Cox regression model for late mortality after aortic valve replacement in young adult patients. When constructing a Cox regression model (or

Table 1. Road map for the use of statistical tests.

	Type of Data			
Goal	**Measurement (from Gaussian Population)**	**Rank, Score, or Measurement (from Non-Gaussian Population)**	**Binomial (Two Possible Outcomes)**	**Survival Time**
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t* test	Wilcoxon test	Chi-square or Binomial test	
Compare two unpaired groups	Unpaired *t* test	Mann–Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel–Haenszel
Compare two paired groups	Paired *t* test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression
Compare three or more unmatched groups	One-way ANOVA	Kruskal–Wallis test	Chi-square test	Cox proportional hazard regression
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q	Conditional proportional hazards regression
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression	Simple logistic regression	Cox proportional hazard regression
Predict value from several measured or binomial variables	Multiple linear regression or Multiple nonlinear regression		Multiple logistic regression	Cox proportional hazard regression

a simple logistic regression model) one first asks the question: which variables may be of clinical importance in determining patient outcome; in this example: which factors may be of influence on late survival. Once these factors have been pre-specified, the first thing to do is to build a univariable model: a model that contains the outcome measure (in this case late mortality occurrence) and one of the potential determinants of outcome that have been pre-specified. In the example it is shown that in the univariable model the type of valve prosthesis that was implanted may potentially be of influence on late mortality, specifically patients with a mechanical prosthesis have a higher mortality rate compared to patients with an autograft valve. After all potential determinants have been tested in a univariable model, one can then start building a multivariable model: a model that puts all univariable statistical significant factors together, and assesses whether these factors are still of importance in predicting late outcome when they are corrected for each other's presence. It is important that before a multivariable model is built, the correlations between the potential predictors are tested and are small; for example: aortic cross-clamp time and cardiopulmonary bypass time are usually highly correlated and one wishes to avoid putting both factors -that express approximately the same issue- together into a multivariable model. In the example in Table 2 it is shown that although in the univariable model patients with a mechanical prosthesis are at increased risk for late mortality, once the multivariable model is constructed, mechanical prosthesis is no longer a significant factor. This can be explained by the fact that patients who receive a mechanical prosthesis more often have a preoperative impaired renal function and left ventricular function, more often require concomitant mitral valve surgery and more often had prior aortic valve surgery. This example illustrates that a complex statistical model can be used to very elegantly explain our complex clinical practice in a simple fashion.

Another example of 21st century statistical tools: In case of comparison between non-randomized groups of patients, statistical methods to match patient populations are available using balancing or more specifically propensity scores that calculate the propensity of each patient belonging to a particular group, and can subsequently be used for direct matching, stratification, or can be forced

Table 2. Risk factors for late mortality. Reproduced with permission from Ref. [13].

Table IV: Risk factors for late mortality.

Risk factor	Univariate analysis model			Multivariate analysis model		
	HR 95%	CI	*p*-value	HR	95% CI	*p*-value
Preoperative impaired renal funcation [1]	1.003	(1.002–1.005)	<0.001	1.004	(1.002–1.006)	<0.001
Preoerative left ventricular function	5.6	(2.6–12.2)	<0.001	5.1	(2.2–11.6)	<0.001
Concomitant mitral valve surgery	3.6	(1.6–8.1)	0.002	3.0	(1.3–7.1)	0.01
Prior aortic valve surgery	3.0	(1.4–6.1)	0.003	3.7	(1.7–7.7)	0.001
Age [2]	1.04	(1.004–1.09)	0.03	1.02	(0.98–1.1)	0.41
Prosthesis type						
Mechanical	8.8	(1.2–65.4)	0.03	0.9	(0.03–1.9)	0.85
Allograft	4.8	(0.6–38.0)	0.14	0.2	(0.4–2.1)	0.18
Autograft (reference group)	1.0	–	–	1.0	–	–

[1] Renal function was analyzed as a continuous variable; the HR represents the increase in risk per additional grade of creatinine.
[2] Age was analyzed as a continuous variable; the HR represents the increase in risk per additional year of age.
HR: Hazard ratio, with 95% confidence intervals (CI).

into a multivariable outcome model. Eugene Blackstone wrote a landmark paper on the subject, entitled "Comparing Apples and Oranges", that is an absolute must-read for all researchers who wish to employ this statistical method [7]. Dr Blackstone warns us to keep in mind that by using this statistical technique we are attempting to compare groups that may potentially not be comparable. He also very nicely explains step-by-step how to construct a propensity score and how to use this score to make predictions about outcome. An example of a study that employs propensity score matching is a recent publication by Mokhles et al. in *Circulation* [7], in which it is attempted to match a patient population of Ross patients with a patient population of mechanical aortic valve recipients, in order to compare survival between the two groups. Patient characteristics differed greatly between the two populations, but in Fig. 2 it is illustrated in a Love plot how by means of propensity score matching it is possible to create two comparable patient populations from two very differing patient populations [8]. In situations where it is not possible of feasible to randomize patients, propensity scores provide a useful tool to compare different differing populations. Although far from perfect [7], they help us paint a clearer picture.

Microsimulation is another more complex statistical method that allows one to estimate for an individual patient lifetime event occurrence and outcome of comparable patients by simulating the postoperative remaining lives of ten-thousands of virtual patients with similar predefined characteristics, based on primary datasets or meta-analyses of outcome after a particular intervention. The fundamentals of microsimulation have been detailed previously, and are useful reading for those who are interested to employ this methodology [9]. Also, an example of a microsimulation model that was built to predict age- and gender-specific patient outcome after aortic valve replacement with different valve prostheses is available as freeware at www.cardiothoracicresearch.nl. Fig. 3 illustrates the use of microsimulation to predict age- and gender-specific event occurrence in patients after aortic valve replacement with either a stented bioprosthesis or a mechanical prosthesis [10]. By using the information from a large dataset of aortic valve replacement (AVR) patients in a microsimulation model it is possible to calculate for patients of a specific age and sex the occurrence of major valve-related complications during the reminder of life. It is evident from Fig. 3 that with increasing patient age the lifetime risk of bleeding increases when a mechanical valve is implanted, while the lifetime risk of a reoperation after bioprosthetic valve implantation decreases. This may be a useful tool for individual counseling or determining optimal treatment strategy in a particular group of patients.

CONCLUSIONS AND FUTURE DIRECTIONS

While modeling, we should realize that we are chasing a moving target: the field of cardiovascular interventions is evolving rapidly with newer less invasive treatment methods that are applied to a

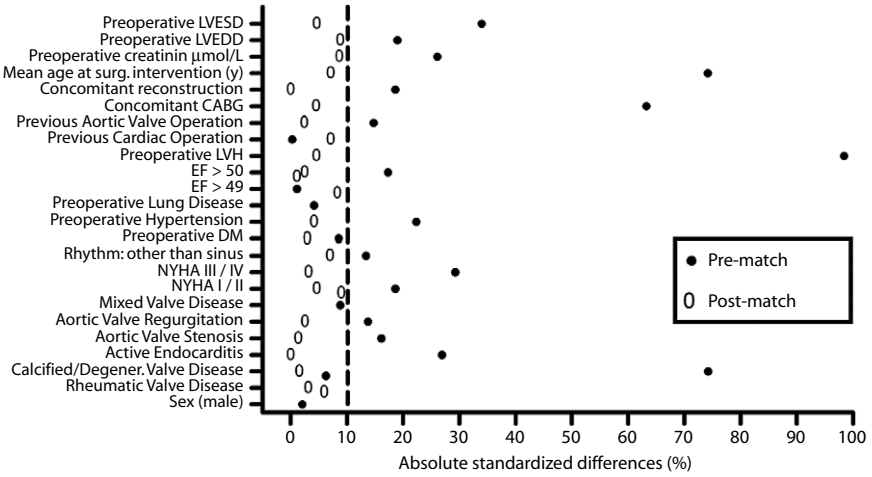

Figure. 2 **Love plots for absolute standardized differences for baseline covariates between patients with the Ross procedure, before and after propensity score matching. DM indicates diabetes mellitus; EF, left ventricular ejection fraction; CABG, coronary artery bypass grafting; MV, mitral valve; and LVEDD, left ventricular end-diastolic diameter; LVESD, left ventricular end-systolic diameter; LVH, left ventricular hypertrophy; and NYHA, New York Heart Association. Reproduced with permission from Ref. [8].**

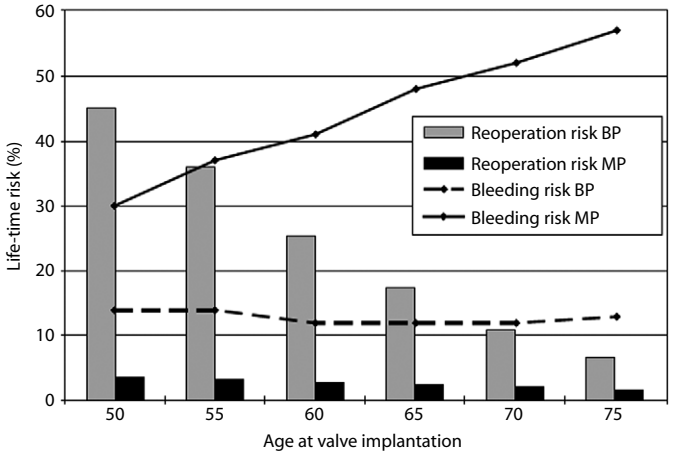

Figure. 3 **Lifetime risks of reoperation and bleeding after AVR with mechanical and bioprostheses. BP, bioprosthesis; MP, mechanical prosthesis. Reproduced with permission from Ref. [10].**

changing patient population of more and more elderly patients with multiple co-morbidities, with mortality risks that continue to decrease. Decreased mortality is a blessing for our patients, but represents a growing challenge for statisticians to accurately predict early outcome based on the characteristics of our patients. We need to rebalance our outcome analysis efforts and besides looking for new explanatory variables that may help us build better models and surrogate outcomes that attempt to reflect quality, also focus on forensic analyses through critical appraisal of the journey of each patient through our hospitals, as Dr de Leval nicely describes in his reflections on outcome analyses [12]. In this respect we can learn a lot from the aviation industry where checklists have proven to be invaluable.

In order to keep up with all innovations in statistical outcome analyses it is highly recommended for clinicians to find in their institution a statistician or statistical department to work with, in order to fully benefit from these advances and optimize statistical modeling of cardiovascular patient data. To quote Albert Einstein one more time:

> "*Any intelligent fool can make things bigger and more complex. It takes a touch of genius — and a lot of courage to move in the opposite direction*".

When we want to find the answer to our carefully defined research question we often need more advanced statistical tools than we can handle as a clinician, and therefore the help of statisticians is essential. On the other hand, as clinicians we need to show leadership when it comes to translating the answer that is constructed using statistics back to clinical practice, from complexity to simplicity. Statistics after all, provide merely a tool to answer clinical questions, and the utility of the statistical methods used depends highly on our ability to translate the answer back to the clinical situation.

References

[1] Roubik DW. Competitive interactions between neotropical pollinators and africanized honey bees. Science. 1978;201:1030–1032.
[2] Hazen RM. Curve-fitting. Science. 1978;202:823.
[3] Stanbrook MB, Austin PC and Redelmeier DA. Acronym-named randomized trials in medicine–the ART in medicine study. The New England Journal of Medicine. 2006;355:101–102.
[4] Diamond GA. Right answer, wrong question: on the clinical relevance of the cardiovascular history. Circulation. 2011;124:2377–2379.
[5] Cheng VY, Berman DS, Rozanski A, Dunning AM, Achenbach S, Al-Mallah M, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow BJ, Delago A, Gomez M, Hadamitzky M, Hausleiter J, Karlsberg RP, Kaufmann P, Lin FY, Maffei E, Raff GL, Villines TC, Shaw LJ and Min JK. Performance of the traditional age, sex, and angina typicality-based approach for estimating pretest probability of angiographically significant coronary artery disease in patients undergoing coronary computed tomographic angiography: results from the multinational coronary CT angiography evaluation for clinical outcomes: an international multicenter registry (CONFIRM). Circulation. 2011;124:2423–2432.
[6] Horton R. Surgical research or comic opera: questions, but few answers. Lancet. 1996;347:984–985.
[7] Blackstone EH. Comparing apples and oranges. The Journal of Thoracic and Cardiovascular Surgery. 2002;123:8–15.
[8] Mokhles MM, Kortke H, Stierle U, Wagner O, Charitos EI, Bogers AJ, Gummert J, Sievers HH and Takkenberg JJ. Survival comparison of the Ross procedure and mechanical valve replacement with optimal self-management anticoagulation therapy: propensity-matched cohort study. Circulation. 2011;123:31–38.
[9] Takkenberg JJ, Puvimanasinghe JP and Grunkemeier GL. Simulation models to predict outcome after aortic valve replacement. The Annals of Thoracic Surgery. 2003;75:1372–1376.
[10] van Geldorp MW, Eric Jamieson WR, Kappetein AP, Ye J, Fradet GJ, Eijkemans MJ, Grunkemeier GL, Bogers AJ and Takkenberg JJ. Patient outcome after aortic valve replacement with a mechanical or biological prosthesis: weighing lifetime anticoagulant-related event risk against reoperation risk. The Journal of Thoracic and Cardiovascular Surgery. 2009;137:881–886. 886e881-885.
[11] Carabello B. Assessment of the patient with valvular heart disease: an integrative approach. Aswan Heart Centre: Science and Practice Series. 2011;15:http://dx.doi.org/10.5339/ahcsps.2011.15.
[12] de Leval M. Reflections on outcome analyses: introducing the concept of near misses. Aswan Heart Centre: Science and Practice Series. 2011;12:http://dx.doi.org/10.5339/ahcsps.2011.12.
[13] Klieverik LMA, Noorlander M, Takkenberg JJM, Kappetein P, Bekkers JA, van Herwerden LA and Bogers AJJC. Outcome after aortic valve replacement in young adults: is patient profile more important than prosthesis type? Journal of Heart Valve Disease. 2006;15:4.