Section 1 Surgery and Critical Care

Chapter 1

Patient outcome following major surgery

Elisa Kam and Rupert Pearse

Introduction

An enormous volume of surgical procedures are performed worldwide each year, particularly in high income nations. While the overall mortality after surgery is relatively low, these figures hide a subpopulation of patients who have much worse outcomes. Given the expansion of the volume of surgery year on year and an increasing tendency to offer surgical treatments to older and high-risk patients, prevention of postoperative deaths has become an important international public health issue. In this chapter, we will review the current knowledge on the mortality and morbidity in patients who undergo surgery, the tools we currently possess to identify the high-risk patient, and the direction of future research.

Surgical outcomes

The need to understand how complications occur and why some patients die as a result highlights the need for robust audit data. Recognition of the importance of audit data is not a new concept. Florence Nightingale described the use of a standard format to report deaths after surgery as early as 1859, and it was her pioneering use of piecharts to illustrate that the majority of deaths in British army hospitals during the Crimean War were due to poor sanitation that helped persuade the British government to improve hygiene in hospitals (Figure 1.1). John Snow (1815–1858), the physician renowned for administering the novel inhalational agent chloroform to Queen Victoria, was the forefather of modern epidemiology and has been credited with promoting quality assessment among his anesthetic colleagues. Leading Boston surgeon Ernest Codman (1869–1940) was one of the earliest to monitor the outcomes of all his surgical patients and, although some of his plans to evaluate the competency of surgeons proved unpopular among his colleagues, his drive for hospital improvement through monitoring outcomes led him to be one of the founders of the American College of Surgeons (ACS).

In more recent times, the voluntary reporting of surgical deaths for peer review has been a popular method of audit. Examples include the Australian and New Zealand Audit of Surgical Mortality (ANZASM), and the Scottish Audit of Surgical Mortality (SASM), which examine all specialties with the exceptions of cardiac, thoracic, and obstetric surgery. However, while these audits report the absolute numbers of deaths of patients who have undergone an operation during a hospital admission, they are limited by the fact that they do not place these deaths in the context of the volume of surgical procedures performed. Recognizing that improving surgical outcomes requires cooperation between all health care professionals involved in the care of surgical patients, a joint initiative began in the UK in the 1980s, now known as the National Enquiry into Patient Outcome and Death (NCEPOD). Through the voluntary return of questionnaires into perioperative deaths and peer review, reports have been published since 1987 and have led to recommendations to improve practice. Notable examples are the creation of operating lists led by senior anesthetists and surgeons during the day for potentially sicker patients undergoing emergency surgery, and introduction of regular departmental morbidity and mortality meetings. The effect of these changes on outcome, however, is less easy to quantify, since the annual number of deaths identified has changed little between 1989 and 2003.

The initiatives of audits and national registries to peer review surgical deaths have enhanced our understanding of the contributing factors and have resulted

Perioperative Hemodynamic Monitoring and Goal Directed Therapy, ed. Maxime Cannesson and Rupert Pearse.
Published by Cambridge University Press.

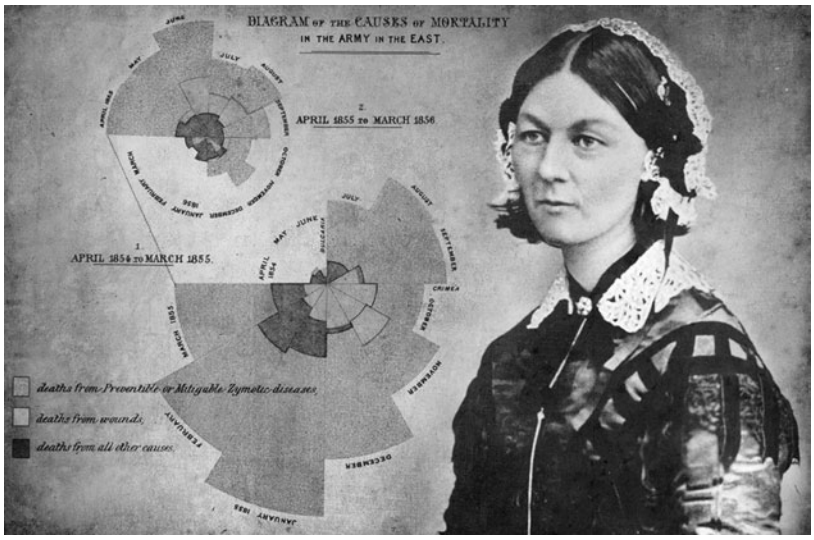

Figure 1.1. Florence Nightingale and the diagram of the cause of mortality in the Army in the East.

in some improvements to clinical practice. However, reporting of absolute numbers of deaths does not yield easily to interpretation. Without a denominator, a mortality rate cannot be calculated. Publications of rates of death and morbidity are much more useful in monitoring standards of surgical and anesthetic care and auditing the effectiveness of planned interventions.

Mortality rates and complication rates

Mortality rates

Our understanding of the epidemiology of surgical outcomes is far from complete. For many surgical specialties, the continuous, prospective collection of accurate data is not yet in place, and mortality rates are estimated from intermittent audits or epidemiological studies. Confusion has arisen from these estimates due to differences in the population studied in terms of geographical boundaries, age exclusions, types of surgery, and the timeframe at which the audit or study was undertaken to obtain the rate of patient deaths.

In the UK, the process of data collection has been pioneered in cardiac surgery, providing robust public audit data on short and medium term mortality not only by hospital, but also by individual surgeons. Risk-adjusted mortality figures are freely available to the public online and these have demonstrated an improvement over time, with a typical hospital mortality rate of 2% or less. Non-cardiac surgery, on the other hand, is less well studied and arguably more important, as the volume of surgery involved is much greater than cardiac surgery and in many cases mortality rates are higher. Although individual registries for specialties such as vascular, bariatric, bowel cancer, and orthopedic surgery report outcomes for certain key procedures, they represent only an individual part of a care pathway and not of the general system of a hospital with a shared perioperative care pathway of standard facilities for preoperative assessment, anesthesia, operating rooms and postanesthetic recovery.

There have been a number of publications of national mortality rates from retrospective analyses of registries and prospective epidemiological studies showing 30-day to 70-month mortality rates for non-cardiac surgery of between 1% and 3%.[1–3] These, however, describe only a small number of health care systems, or parts of a national health care system. In a recent study, investigators have attempted to study a larger number of health care systems. The European Surgical Outcomes Study was an international prospective study of 46 000 adult patients from 28 European countries undergoing non-cardiac surgery over a 1-week period. It showed a higher overall 60-day mortality of 4%.[4] Although the overall mortality in non-cardiac surgery appears relatively low, mortality may exceed 12% in older patients undergoing emergency surgery (Figure 1.2). A small group of high-risk patients has been shown to be responsible for 84% of deaths and significantly longer hospital stays, despite making up only 12.5% of hospital admissions for surgery.[1] The significance of identifying and caring for this group of patients is highlighted in the next section.

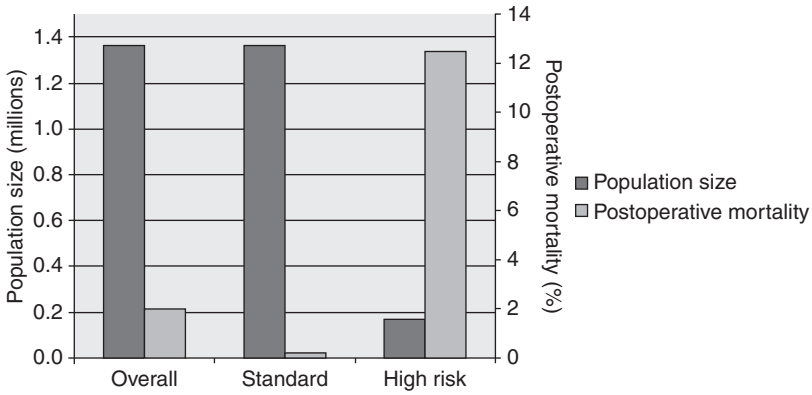

Figure 1.2. Mortality rates for different surgical populations in the UK.

Complication rates

A precise estimate of perioperative complications is difficult to provide, but they may occur following 15% to 27% of all surgical procedures.[5,6] The wide range in complication rates reflects variable reporting and also the large number of possible anesthetic or surgical complications covering many organ systems, including cardiovascular, pulmonary, renal, hematological, and gastrointestinal, as well as infections. Models have been devised to help classify these broad categories of surgical complications. Clavien proposed a model in the 1990s,[7] which has since been updated and validated in a large cohort of patients by an international survey to allow grading of severity of postoperative complications, regardless of the initial surgery.[3] Another model is the Postoperative Morbidity Survey (POMS), which is a validated questionnaire developed to record postoperative complications in non-cardiac surgery.

Collecting information on postoperative complications is important not only for audit purposes, but there has also been an increasing recognition that developing complications increases a patient's risk of death. Patients who develop complications but survive may suffer a substantial reduction in functional independence and long-term survival. Analysis of data from the United States' National Surgical Quality Improvement Program (NSQIP) showed that not only did the occurrence of 30-day postoperative complication reduce patient survival by 69%, but it was more important than preoperative and intraoperative factors in determining survival after major non-cardiac surgery.[3] Another large North American study showed that mortality in an unselected population of surgical patients doubled from 2% to 4% after surgery and by 1 year, 47% of surviving patients had been readmitted to hospital.

Why are mortality and morbidity so high?

Advances in surgical techniques, training, and increased subspecialization have led to significant improvements in care. Concurrently, mortality directly attributable to anesthesia has declined steeply. Despite such improvements in patient treatments during surgery, mortality after surgery has not declined. There is increasing recognition that the care of the patient after an operation is equally important in determining outcomes, and growing concern that it is the quality of this postoperative care that is not of a high enough standard. More surgery is being performed on patients with higher risk as a result of an aging world population with more co-morbid disease, as well as more operations in younger patients who have a higher illness burden. It is estimated that there is a subgroup of high-risk patients that accounts for 80% of all postoperative deaths.[1,8] Epidemiological data suggest that clinicians often fail to identify these high-risk patients preoperatively in order to plan appropriate perioperative care. It is estimated that 170 000 of high-risk patients will undergo non-cardiac surgery in the UK and that 60% of these patients will develop complications after non-cardiac surgery, leading to over 25 000 deaths.[1,8]

There is evidence that critical care-based cardiorespiratory interventions can improve outcomes in high-risk patients. Cardiac surgery in traditionally high-risk patients will routinely admit the majority of its patients to critical care postoperatively. However, critical care provision is low for patients undergoing non-cardiac surgery. Unplanned admissions to critical care are associated with higher mortality rates than planned admissions, yet only 5% of patients undergoing non-cardiac surgery have a planned admission to critical

care.[9] A review of Medicare data reveals that the differences in mortality between hospitals are related to the ability of a hospital to effectively rescue patients from complications.[10] This suggests a failure to recognize the sick and high-risk patient and perhaps the lack of availability of critical care resources.

Risk assessment

Identifying the patients who are most likely to suffer postoperative complications or mortality allows informed decisions on whether to operate and to help target postoperative care and critical care provision for these patients. The majority of patients are evaluated prior to surgery solely according to the physician's assessment of clinical history, physiology, and extent of surgery. Several tools have been developed to assist the clinician in predicting the response of a patient to the tissue injury induced inflammatory state of surgery, but many of these are not yet in routine use due to cost, ease of use and a developing evidence base. These tools include risk scores, serum biomarkers, and assessment of functional capacity.

Risk scores

General scores are used to estimate population risk. One of the earliest systems proposed is the American Society of Anesthesiologists (ASA) classification,[11] which stratifies a patient's ability to withstand surgery into one of five classes depending on the presence and severity of co-morbid disease. Although initially developed as a tool for audit and research, the individual ASA classes may be used as predictors of mortality, while the rate of postoperative morbidity varies with class. The ASA system is popular because it is easy to use, but scoring can be subjective and it does not allow consideration of individual specific information or the type of surgery. Hence, the system has poor sensitivity and specificity when used to assess the risk in an individual patient.

The Acute Physiology and Chronic Health Evaluation (APACHE) scores were developed in the 1980s for use in critical care. Of the four versions, the APACHE II score is the most validated for use in preoperative risk prediction. It is based on 12 physiological variables, with additional points for age and chronic health.[12] While the type of surgery is not accounted for, APACHE II can provide an individualized risk of morbidity and mortality. It performs better in predicting outcome than the ASA classification[13] and may be used to predict severity of surgical complications. However, the requirement to measure all variables over the first 24 hours of critical care stay before an operation is a barrier to the regular use of this score.

Goldman and Lee have produced well-validated scoring systems to predict the likelihood of cardiac complications after non-cardiac surgery,[14,15] and Arozullah has provided a model for predicting postoperative respiratory failure.[16] However, these focus on a single organ system and cannot make assessment of the severity of each contributing factor.

The Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity (POSSUM) score was designed for use in preoperative risk prediction, whilst taking into consideration both individual physiological risk and the type of surgery performed.[6] This scoring system uses 12 physiological and six operative variables to predict mortality and morbidity via two separate equations. POSSUM may overestimate or underestimate risk in specific populations, but it remains the most validated and used scoring system for non-cardiac surgery.

In cardiac surgery, the European System for Cardiac Operative Risk Evaluation (EuroSCORE) is one of the most validated for hospital and long-term mortality.[17] It is calculated using clinical data either in an additive or logistic calculation, the former being easier to derive but less accurate in high-risk patients. It is widely used in research and audit, but caution has been recommended for comparisons and for surgeons with different case mixes.

Biomarkers

There has been a growing interest in biochemical markers with the goal of finding an inexpensive biochemical test that either alone or in combination with existing clinical tools can improve the accuracy of perioperative risk prediction. Several systemic reviews and meta-analyses and observational studies[18–20] suggest that elevated serum concentrations of high-sensitivity C-reactive protein (hs CRP) and N-terminal pro-B-type natriuretic peptide (NT pro-BNP) prior to surgery may be independent predictors of adverse cardiac events in medium or short term following major non-cardiac surgery. Moreover, these preoperative values can be used to prognosticate cardiac complications and mortality after high-risk surgery. Serum concentrations of troponin taken during the postoperative period have also been shown to be a strong independent predictor of short-term mortality in non-cardiac surgery.

Measurement of cardiac troponin levels for the first 3 days after surgery may substantially improve the accuracy of 30-day mortality risk stratification compared with assessment limited to preoperative risk factors.[21] Markers for neurological damage such as S100B, Tau, and the enzyme neurone-specific enolase have been assessed in cardiac and non-cardiac surgery, but results are conflicting.[22]

Functional capacity

Low exercise tolerance is associated with poor outcomes.[23] Preoperative assessment of functional capacity aims to predict an individual's ability to increase oxygen delivery during the perioperative period. Tests such as echocardiography and spirometry are useful but limited as they are performed at rest. The European Society of Cardiology (ESC) and the American College of Cardiology/American Heart Association (ACC/AHA) guidelines recommend using the metabolic equivalent task (MET) as an estimate of functional capacity. One MET is the metabolic requirement of an activity such as walking around indoors or doing light housework and is equivalent to 3.5 ml O_2/kg. The threshold of acceptable functional capacity is given as four METs, which is equivalent to climbing a flight of stairs. It is important to recognize there are inaccuracies in this method of estimation because the definition of MET is derived from the measurement of resting oxygen consumption from a single 70 kg, 40-year-old man. Thus an accurate assessment of functional capacity requires the knowledge of an individual's resting oxygen uptake, as well as reliable reporting of functional activity from the patient.

Cardiopulmonary exercise testing (CPET) is the gold standard method for assessing an individual's functional capacity by measuring oxygen uptake and carbon dioxide elimination while performing incremental, symptom-limited physical exercise up to the patient's maximal level. Incorporating ECG monitoring, it provides an integrated look at both a patient's cardiac and respiratory function during exercise. The main values of interest are the body's peak oxygen consumption (VO_2 peak) and anaerobic threshold (AT). Patients are classified as being at increased risk if VO_2 peak is less than 15 ml O_2/kg and AT is less than 11 ml O_2/kg/min. [23] CPET testing has a good predictive value for postoperative complications in pulmonary resection surgery, and there is increasing evidence of benefit in predicting morbidity and mortality in general surgery. However, CPET requires investment into costly equipment and skilled personnel to perform and interpret the tests, and for some surgical subspecialties there are still doubts over the evidence base, and this has prevented its routine use.

Large international trials are planned to define the optimal approach to evaluate risk assessment prior to surgery. It can be envisaged that all patients may be offered initial screening, based on simple factors such as age, type of surgery, serum biomarkers, and clinical risk scores. Low-risk patients could be offered early surgery following assessment in the community, while complex patients could be offered more sophisticated tests and detailed assessment by a physician. This would improve informed discussions with patients and allow individualized treatment plans with optimal use of postoperative critical care resources.

System-wide strategies to improve surgical outcomes

The design of health care systems significantly impacts on a hospital's ability to detect and manage postoperative adverse events and hence clinical outcomes. Outcome measures are increasingly used to underpin quality improvement frameworks and guide purchasing or commissioning of health care services. In the USA, the Centers for Medicare and Medicaid Services now deny reimbursement to hospitals for specific postoperative adverse events including urinary tract infections, pressure ulcers, and surgical site infections.[5] Such financial penalties may help drive quality improvements in other health care systems.

Various targets along the patient care pathway have been identified for patient safety and quality improvement initiatives. Many of these have been described above, including preoperative risk assessment, and joint clinics involving surgeons, anesthetists, and physicians allowing effective decision making and better communication with community health care teams. Similarly, systems should facilitate effective treatment plans for patients with a delayed recovery after hospital discharge, allowing a prompt return to hospital for review by the surgical team. Important structural factors include availability of critical care beds, staffing levels, and working patterns which influence the specialization of staff involved in the care of the surgical patient. These factors affect the ability of a system to ensure optimal treatment at the time of surgery and to promptly identify and treat those patients who later deteriorate. While there is no

direct evidence of the benefits of these systems, they are known to exist more commonly in centers treating large volumes of patients. A clear association exists between hospital volume and clinical outcomes for many complex surgical procedures and high-risk patients.[24] Low-risk patients, however, have been shown to have comparable outcomes in both low-volume and high-volume centers. Hence, it is important to be able to risk-stratify a patient before selecting the most appropriate hospital for an elective operation.

Effective hospital clinical governance is key to delivering high-quality care. This incorporates complete and accurate data collection, internal audit, benchmarking against defined quality standards, and transparent publication of results. The USA has led the way in large-scale quality improvement projects. In the 1990s, data were prospectively collected for major operations in some Veteran Affairs hospitals and used to develop risk-adjusted models for 30-day morbidity and mortality.[25] Hospitals with lower morbidity and mortality were used as a standard against which adjustable factors in individual hospitals with worse outcomes could be identified. Use of data in this way by the National Surgical Quality Improvement Program (NSQIP) resulted in a 45% decrease in morbidity and a 27% reduction in mortality across all the Veteran Affairs hospitals. By 2008, and with funding from the American College of Surgeons (ACS), ACS-NSQIP was expanded to 198 hospitals across the USA. In the UK and in many other countries, clinical benchmarking is only available for selected procedures, and this represents only a small percentage of overall surgical volume. There is therefore an urgent need for investment in effective and widespread audit of surgical care and outcomes.

Conclusions

Rates of mortality and complications postsurgery have been difficult to collect for non-cardiac surgery, with estimates derived from registries and national databases. However, until mortality and morbidity tables for individual hospitals and surgeons are routinely published, hospital managers and clinicians will be limited in their attempts to improve outcomes for their patients. Financial constraints on health services around the world have led to remuneration based on outcome, and this may yet be the driving force for investment into clinical governance and quality improvement programs. At present, we are poor at identifying high-risk patients who are more likely to suffer adverse events. Preoperative scoring has the potential to ensure better informed consent and patient/procedural selection. The possibility of individualized risk prediction based on an individual's physiological response to stress is an exciting area, with the possibility of high predictive value and better use of critical resources to improve patient care

References

1. Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Crit Care* 2006;**10(3)**:R81.
2. Yu PC, Calderaro D, Gualandro DM, et al. Non-cardiac surgery in developing countries: epidemiological aspects and economical opportunities – the case of Brazil. *PLoS ONE* 2010;**5(5)**:e10607.
3. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg* 2004;**240(2)**:205–13.
4. Pearse R, Moreno RP, Bauer P, et al. Mortality after surgery in Europe – authors' reply. *Lancet* 2013;**381(9864)**:370–1.
5. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. *N Engl J Med* 2009;**361(14)**:1368–75.
6. Bennett-Guerrero E, Hyam JA, Shaefi S, et al. Comparison of P-POSSUM risk-adjusted mortality rates after surgery between patients in the USA and the UK. *Br J Surg* 2003;**90(12)**:1593–8.
7. Clavien PAP, Sanabria JRJ, Strasberg SMS. Proposed classification of complications of surgery with examples of utility in cholecystectomy. *Surgery* 1992;**111(5)**:518–26.
8. Jhanji S, Thomas B, Ely A, et al. Mortality and utilisation of critical care resources amongst high-risk surgical patients in a large NHS trust. *Anaesthesia* 2008;**63(7)**:695–700.
9. Pearse RM, Moreno RP, Bauer P, et al. Mortality after surgery in Europe: a 7 day cohort study. *Lancet* 2012;**380(9847)**:1059–65.
10. Ghaferi AAA, Birkmeyer JDJ, Dimick JBJ. Hospital volume and failure to rescue with high-risk surgery. *Med Care* 2011;**49(12)**:1076–81.
11. Saklad M. Grading of patients for surgical procedures. *Anesthesiology* 1941;**2(3)**:281.
12. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II:

a severity of disease classification system. *Crit Care Med* 1985;**13 (10)**:818.

13. Goffi L, Saba V, Ghiselli R, et al. Preoperative APACHE II and ASA scores in patients having major general surgical operations: prognostic value and potential clinical applications. *Eur J Surg* 1999;**165 (8)**:730–5.
14. Goldman L, Caldera DL, Nussbaum SR, et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. *N Engl J Med* 1977;**297(16)**:845–50.
15. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999;**100 (10)**:1043–9.
16. Arozullah AM, Khuri SF, Henderson WG, Daley J. Development and validation of a multifactorial risk index for predicting postoperative pneumonia after major noncardiac surgery. *Ann Intern Med* 2001;**135(10)**:847–57.
17. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation* 1989;**79(6 Pt 2)**:I3–12.
18. Karthikeyan G, Moncur RA, Levine O, et al. Is a pre-operative brain natriuretic peptide or N-terminal pro–B-type natriuretic peptide measurement an independent predictor of adverse cardiovascular outcomes within 30 days of noncardiac surgery? A systematic review and meta-analysis of observational studies. *J Am Coll Cardiol* 2009;**54 (17)**:1599–606.
19. Choi J-H, Cho DK, Song Y-B, et al. Preoperative NT-proBNP and CRP predict perioperative major cardiovascular events in non-cardiac surgery. *Heart* 2009;**96(1)**:56–62.
20. Goei D, Hoeks SE, Boersma E, et al. Incremental value of high-sensitivity C-reactive protein and N-terminal pro-B-type natriuretic peptide for the prediction of postoperative cardiac events in noncardiac vascular surgery patients. *Coron Artery Dis* 2009;**20 (3)**:219–24.
21. Devereaux PJ, Chan MT, Alonso-Coello P, et al. Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. *JAMA* 2012;**307 (21)**:2295–304.
22. Cata JP, Abdelmalak B, Farag E. Neurological biomarkers in the perioperative period. *Br J Anaesth* 2011;**107(6)**:844–58.
23. Older P, Hall A, Hader R. Cardiopulmonary exercise testing as a screening test for perioperative management of major surgery in the elderly. *Chest* 1999;**116(2)**:355–62.
24. Birkmeyer JDJ, Siewers AEA, Finlayson EVAE, et al. Hospital volume and surgical mortality in the United States. *N Engl J Med* 2002;**346(15)**:1128–37.
25. Khuri SF, Daley J, Henderson W, et al. The National Veterans Administration Surgical Risk Study: risk adjustment for the comparative assessment of the quality of surgical care. *J Am Coll Surg* 1995;**180(5)**:519–31.

Section 1 Surgery and Critical Care

Chapter 2

Statistical methods in hemodynamic research

Yannick Le Manach and Gary Colins

Introduction

One of the most common research questions in hemodynamics concerns the comparison of measurement devices, and the determination of specific values in observed parameters. In this chapter, we will provide an overview of the statistical methods used to address them. The cardiac output measurements devices evaluation and the determination of threshold for preload dependency parameters will be used for illustration.

To compare two methods

Hemodynamic research often concerns comparing two methods of measurement. Typically, a new measurement method is being compared with an established method (often referred to as the "gold standard"), to determine whether the two methods can be used interchangeably, or if the new method can replace the established one. Interchangeability refers to the ability of one measurement device to be replaced by another one without affecting clinical interpretation of the observed values. Cardiac output measurement devices are among the most studied and compared devices. They constitute a good support to discuss the methodology, and most of the accumulated knowledge on them can be used to compare other sorts of devices.

Correlations

One of the most commonly used approaches to compare two methods of measurement is to calculate the correlation between the two methods. However, the correlation between two methods of measurement is uninformative and does not actually assess the agreement between the two methods. Correlation measures the strength of linear association between two continuous measurements. A perfect correlation ($r = 1$) occurs when the measurements of the two methods lie on any straight line. However, if we consider the situation whereby the new measurement method outputs a measurement exactly twice that of the standard method, then the correlation is still 1, yet clearly the two methods are not interchangeable. A change of scale in one of the two measurement methods does not affect the correlation, yet it clearly affects the agreement (or lack of). The interpretation of the correlation coefficient between two methods includes some limitations. When a very high correlation (e.g., $r = 0.95$ with $P < 0.001$) is observed, the probability to reject the null hypothesis (e.g., no linear relationship between the two sets of measurements) is very small and we can safely conclude that measurements by both devices are related. However, this high correlation does not mean that the two methods are interchangeable. In fact, as noted earlier, a change in scale of measurement will not affect these correlations, but it certainly affects the agreement and the potential clinical use of the new device. Furthermore, correlation depends on the range of the cardiac outputs in the sample. If the range of observed values is wide, the correlation will be greater than if it is narrow. Since researchers usually try to compare two methods over the whole range of cardiac output values typically encountered, a high correlation is almost guaranteed. Finally, a test of significance may show that the two methods are related, but it would be remarkable if two methods designed to measure cardiac output were not related.[1]

Scatter plots remain useful because they depict the relationship (linear or not) between the devices, as well as the difference in the measure for each level of cardiac output. However, we have to assume that

Perioperative Hemodynamic Monitoring and Goal Directed Therapy, ed. Maxime Cannesson and Rupert Pearse.
Published by Cambridge University Press.

there is no simple estimator as soon as the statistical relationship is not linear. For this purpose, correlation, and more specifically correlation plots, might be considered in the reporting of studies aiming to compare two cardiac output measurement devices. Nevertheless, non-linear relationships are also identifiable in limit of agreements plots, and the researchers may consider that reporting the two is probably not necessary.

Although widely used in the literature, correlation analyses are not required in most of the cases. Although low correlation reflects low agreements between devices, high correlation doesn't necessarily refer to interchangeability. Correlation plots provide some details about the statistical relationship of interest, but this can be easily read on other analyses. If the inclusion of such analysis in research reports remains disputable, conclusions done only on correlation analyses are not acceptable.

Bland and Altman's graphical representation

In response to the widespread and inappropriate use of the correlation coefficient to assessment method agreement, statisticians Martin Bland and Douglas Altman proposed an alternative approach based on graphical techniques (known as the Bland–Altman plot).[2] This landmark *Lancet* paper by Bland and Altman has to date been cited on more than 22 000 occasions, illustrating its importance in medical research. The primary application of the Bland and Altman agreement plot is in the comparison of two clinical devices that contain error in their measurement. The aim is to determine how much the two measurement methods are likely to differ. If this difference is sufficiently small not to cause problems in clinical interpretation, then it can be considered as a candidate to replace the old method or to be used interchangeably.

The Bland and Altman plot allows us to investigate the existence of any systematic difference between the measurements. It is a plot of the difference against the mean of the two measurements. The mean difference is the estimated bias. If the mean value of the difference differs significantly from 0, this indicates the presence of fixed bias. If the differences lie between the 95% limits of agreement (mean ± 1.96 SD), then they are deemed not important; the two methods may be used interchangeably. Bland and Altman plots have also been used to investigate any possible relationship of the discrepancies between the measurements and the true value (i.e., proportional bias). The existence of proportional bias indicates that the methods do not agree equally through the range of measurements (i.e., the limits of agreement will depend on the actual measurement). To evaluate this relationship formally, the difference between the methods should be regressed on the average of both methods. When a relationship between the differences and the true value was identified (i.e., a significant slope of the regression line), regression-based 95% limits of agreement should be provided

The presentation of the 95% limits of agreement is for visual judgment of how well two methods of measurement agree. The smaller the range between these two limits the better the agreement. The question of how small is small depends on the clinical context: would a difference between measurement methods as extreme as that described by the 95% limits of agreement meaningfully affect the interpretation of the results?

In the case of cardiac output devices comparison, it has been suggested that the limits of agreement between two methods should approach the precision of the older reference method before accepting the newer technique.[3] Commercial thermodilution devices (i.e., the gold standard for cardiac output measurement) are recognized to have a minimal difference of 12 to 15% (average, 13%) between measurements of cardiac output.[4] Thirteen percent has thus been described as the 95% limits of agreement in studies focusing on cardiac output devices comparison. Nonetheless, a 13% difference should be interpreted differently according to the absolute value of the cardiac output, as this 13% was obtained by averaging three measures, whereas a 22% difference was observed when only one measurement was used per determination. Consequently, 13 is not a magic number and researchers should consider higher variability in their reference measurements, particularly when their protocol does not include an averaging of at least three measurements.

It has also been suggested that the results of such evaluation studies should include the mean cardiac output, the bias, and the 95% limits of agreement.[5] Furthermore, it was suggested to report percentages rather than absolute values. The 95% limits of agreement do not include the variability of the reference method (e.g., 13% for thermodilution). Consequently, the observed disagreement is entirely assigned to the new method. Critchley et al.[5] suggested a corrective method to take into account the variability of the reference methods. Using an error gram, they depicted the relationship between the accuracy of the reference

method and the limits of agreement between the new and the reference technique. As an example, they calculated that a limit of agreement of 1.45 L min^{-1} (28.3% of error) represented a clinically relevant disagreement when the reference methods presented variability of 20% and with a mean cardiac output of 5 L min^{-1}. In this setting, they recommended that limits of agreement between the new and the reference technique of up to 30% be accepted.[5]

Although the Critchley et al.[5] demonstration was compelling, the 30% limits of agreement are appropriate only when the variability of the reference method is 20%. This may be the case in most of the studies using averaged thermodilutions as the reference method; however, there is a great risk of rejecting some new methods only because the "reference" method was not as accurate as it should have been (e.g., not averaged thermodilution, alternative method used as reference).

Repeated measurements for each subject are often used in hemodynamic research. When repeated measures data are available, it is desirable to use all the data to compare the two methods. Several alternatives are available for the analysis of repeated measures;[6,7] among them the method described by Bland and Altman appears to be the simplest.[6]

Trend analysis

We described approaches aimed at evaluating the agreement between measures provided by two devices. However, in some clinical settings, these absolute values have limited clinical interest. Instead, the temporal variability of these absolute measures is of interest (referred to as trend analysis). There is no doubt that interchangeable devices would likely present a very high level of agreement in trend analysis. However, some devices presenting a bias in the measurements could be interesting when trends are considered. Repeated measures approaches are not able to evaluate the agreement between trends. The difference between the points of measurement for both methods can be analyzed using a Bland and Altman plot, and absolute or relative variations can be used with this approach. However, absolute variations do not take into account the baseline cardiac output (i.e., mean value before the change) and may not be able to identify clinically relevant disagreements with a wide range of baseline cardiac output. Relative variations do not take basal cardiac output into account; however, the percentage of variation may help the researchers to conclude about the clinical impact.

Critchley et al. have recently introduced a new approach to compare the trends between two cardiac output devices.[8] This new method addresses the magnitude of change between pairs of consecutive readings and the degree of agreement. A circular graph, called a polar plot, is proposed, which requires the changes in measurements given by the two cardiac output measurement devices to be transformed to polar coordinates. The best description of polar plot is given by Critchley:[9]

> *In the polar plot, the ΔCO data are converted to a radial vector where the degree of agreement between the 2 devices becomes the angle between the radial vector and the horizontal axis (i.e., polar axis). If agreement is perfect, the radial vector will lie along the polar axis and the angle is zero. The mean angle from all the radial vectors is the mean polar angle and is the statistic used to measure agreement. It is continuous rather than binomial variable (i.e., agree or disagree). The distance from the center of the polar plot or radius represents the magnitude of ΔCO in the polar vector and is derived from the average of reference and test ΔCO.*

Methodological concerns

Studies evaluating cardiac output measurement devices generally do not specify a clinically meaningful limit of acceptable agreement before the analyses were conducted. Clinical interpretation of the interest of the new device is often done based on the results (i.e., posthoc). This approach is flawed because the width of the 95% limit of agreements confidence interval is mainly determined by the achieved sample size. Although we are dealing with an estimation problem, the estimation of required or desirable sample size is as relevant as it is to inference. Few calculations are needed to demonstrate that the size of confidence interval for the 95% limit of agreement is a function of the standard deviation of the differences between measurements by the two methods and of the sample size.[10] Researchers are thus able to determine the expected width of this confidence of interval. In fact, most of the studies aiming to evaluate cardiac output measurement devices are conducted on small sample sizes (e.g., less than 100 pairs of independent measurements). This impacts considerably on the robustness of the conclusions about the possible interchangeability of the devices. Estimation of the width of confidence interval for the 95% limit of agreement should be calculated a priori and used to determine the number of patients needed to get a robust estimation of the device.