

Current system unlikely to pick up surgeons with above average patient death rates

Performance within 'expected' range not good enough to detect outliers, say researchers

Publishing the patient death rates of individual surgeons in England is unlikely to pick up those whose mortality rates are above average, because the caseload varies so much, concludes the first analysis of its kind published in the online journal **BMJ Open**.

Performance within the 'expected' range is too crude a measure to detect doctors whose practice might be a cause for concern, and is therefore creating a false sense of security, say the researchers.

When the patient death rates for individual surgeons were first published in June 2013, the move was hailed as a major breakthrough in transparency that would drive up standards of care in England.

But the chances of detecting a surgeon whose death rates are worse than the national average is a question of statistical power, say the researchers: in other words, the greater the caseload, the greater the ability to detect worrying trends.

To assess how reliable the available data for individual surgeons are, the researchers reviewed the outcomes for three common high risk procedures—bowel surgery, gullet surgery, and planned aortic aneurysm repair—and three common low risk procedures—hip replacement, bariatric surgery, and thyroid removal.

And they analysed every surgeon's caseload for each of the procedures, all of which were carried out between 2010 and 2014 across England.

They focused in particular on how well these data would be able to detect a surgeon whose patient death rate in hospital or within 30 or 90 days of the patient's discharge was between two and five times higher than the national average.

Unsurprisingly, the higher risk procedures were associated with a higher death rate of between 2.2-4.5% while the lower risk ones were associated with a death rate of 0.07-0.4%.

But caseload was an issue. For example, the average number of bowel surgery operations carried out by individual surgeons was 55 over three years, but ranged from just 3 to 237.

With an average national 90 day death rate of 3%, the national average of 55 cases provides 20% statistical power to detect a mortality rate three times the national average. That means that around 20 out of 100 individual surgeons with an actual death rate of 9% would fall outside the expected range.

But the caseload would have to be more than 200 to provide 90% statistical power of detecting a surgeon whose 90 day mortality rate is three times the national average.

Similar findings emerged for gullet surgery, where the average number of procedures was 23 over a two year period, but ranged from 10 to 81.

Based on national 30 day death rates of 2.4%, the average number of cases would provide less than 20% statistical power to detect a surgeon with a patient death rate four times the national average.

And a caseload of 300 procedures would be needed to provide 80% statistical power to detect a 90 day mortality rate twice as high as the national average over two years.

For low risk procedures, the national average caseload ranged from 48 to 75 per surgeon, meaning that fewer than 20 out of 100 surgeons with an actual mortality rate five times the national average would be picked up.

For hip replacements, for example, an annual caseload of more than 500 cases would be needed to provide 80% statistical power to pick up just one individual with a mortality rate five times the national average.

At these kinds of rates it is unlikely that a surgeon would ever perform enough procedures in his/her entire career for a mortality rate five times the national average to be detected, say the researchers.

“On the basis of these rates and published case volumes, surgeons with mortality rates in excess of that expected are highly unlikely to be detected,” they write. “Performance within an expected mortality rate range cannot therefore be considered reliable evidence of acceptable performance.”

More meaningful outcome measures are required, they say. These could include patient satisfaction, the ease with which routine daily tasks can be performed (functional health status), and other health related quality of life indicators.

And an individual’s performance could be addressed by regular internal appraisal and feedback from multiple sources, they suggest.

Interpreting performance data for individual surgeons has major implications for patient care, the individual practitioner, and their employer, they emphasise

But they conclude: “This analysis demonstrates that, for these common procedures, mortality rates are not a robust method for detecting divergent practice. It is not surprising that the performance of all but one surgeon across all six procedures was found to be acceptable.”