024: SUMMARY MEASURES OF QUALITY OF CLINICAL CARE: MAPPING THE MINEFIELD.

Authors
Reeves D., Roland M., Shekelle P.

Objective:
To explore various methods for deriving summary measures of quality of clinical care, to assess each with respect to the method's statistical properties and the 'clinical meaning' of the underlying assumptions and resulting scores, and to suggest some of the characteristics required of an 'ideal' measure.

Methods:
The quality of the clinical aspects of patient care is frequently assessed through a set of indicators, each of which relates to a distinct element of clinical care. In the case of a diabetes mellitus patient, for example, whether HbA1C levels have been checked regularly, and whether annual fundal examinations have been undertaken. It is normally desirable to derive some kind of 'summary' or 'aggregated' score by combining across indicators to produce an overall score for a particular medical condition, or - for instance - for a general practice as a whole. However, the approach to calculating summary scores represents something of a 'minefield': many different methods can be applied, none of which is definitive and each of which possesses particular advantages and disadvantages, both statistically and in terms of clinical meaning.

The summarising process can be thought of as having two distinct components: first, the algorithm used to combine across indicators; and second, the weights (if any) given to each indicator (to reflect clinical importance, say). Different combinations of algorithm and weights produce different summary scores: for example, some algorithms tend to produce scores dominated by higher frequency indicators; weighting systems differ in the extent to which performance on certain indicators or conditions contribute to the final score. In this way the choice of algorithm and weights has an effect not only on the resulting quality-of-care score, but also on the clinical meaning of that score, and can potentially lead to different conclusions about a practices' performance.

This paper applies four different summary score algorithms and a range of weighting systems to data from 16 GP practices that participated in the Global Quality Assessment Project. The analysis is focused on the effect of different methods on resulting practice-level summary scores. Varying the weighting system has an impact on the scores that is largely determined by the extent to which the individual indicators correlate with one another, and this feature is also used to explore the general sensitivity of the summary scores to changes in the weights.

Results:
Results are presented for each of the four algorithms, comparing the impact of each on practice-level summary quality scores and practice rank-order, under a range of weighting schemes. Data on inter-indicator correlations and weight-sensitivity are also presented. Finally, the statistical and clinical pros and cons of each method are described and discussed.

Conclusions:
The derivation of quality-of-care indicators is a relatively new science, and the development of appropriate and agreed-on methods is still at an early stage. This paper makes a contribution to the debate by illustrating the application of a range of methods to real data; demonstrating the degree to which the different methods alter the final scores and thus the conclusions about quality that stem from these; and revealing the relation between clinical meaning and the choice of algorithm and weights. The investigation also throws some light on the characteristics that are desirable in an 'ideal' quality-of-care measure.