

EXECUTIVE SUMMARY

**REPORT TO THE ONTARIO HOSPITAL REPORT RESEARCH
COLLABORATIVE**

Gwyn Bevan and David Spiegelhalter

Gwyn Bevan is Professor of Management Science at the London School of Economics & Political Science [R.G.Bevan@lse.ac.uk]

David Spiegelhalter is a Senior Scientist in the MRC Biostatistics Unit at Cambridge University
Institute of Public Health
[david.spiegelhalter@mrc-bsu.cam.ac.uk]

August 2006

Report to the Ontario Hospital Report Research Collaborative

Executive Summary

Hospital Reports in Ontario, Scotland and England

- i. The Ontario Hospital Report Research Collaborative (HRRC) is a superb example of a research collaboration that has produced important pioneering *Hospital Reports* in assessing health care performance. There are now *Hospital Reports* on Acute Care, Rehabilitation, Complex Continuing Care (CCC), Emergency Department (ED) and Mental Health (which is under development). *Hospital Reports* are organised in the form of a balanced scorecard across four quadrants: System Integration and Change (SIC), Patient Satisfaction; Clinical Utilization and Outcomes (CUO); Financial Performance and Condition. From 2005, a Women's Health Perspective was added. The information used for each quadrant naturally varies by sector. Performance by institution was colour coded as being 'above average', 'average' or 'below average'. Less detailed information was reported in the recent development of the Women's Health Perspective.
- ii. We compared HRRC's *Hospital Reports* with annual Reports by the Clinical Resource and Audit Group (CRAG) in Scotland and the English system of star ratings. Ontario *Hospital Reports* are far superior to the other two in their well-developed assessment of financial performance and in emphasising coordination across sectors and the community. But Ontario *Hospital Reports* lack information on public health, primary care, waiting times; and assessments of patient and public involvement, clinical audit, and risk management.
- iii. CRAG reports suffered from a number of weaknesses, which meant that they failed to have an impact. Star ratings did result in dramatic improvements in reported performance on the most important targets resulted, but it is unclear to what extent these were undermined by gaming and the neglect of services that were not targeted. We have recommended that HRRC finds out more about the impacts, and improves publicity and dissemination of their *Hospital Reports*; develops methods of aggregation for the target audience (of directors and senior managers); and explores how information can be better used for benchmarking.
- iv. HRRC is an example of a 'soft' approach to reporting on health care performance: this is directed at quality improvement and emphasises identifying and celebrating success (known as 'naming and faming'), is voluntary and not part of government. The system of star ratings is an example of a 'hard' approach: this emphasised accountability with requirements to meet minimum standards, hence the sanctions for failure were much clearer than the rewards for success (known as 'naming and shaming'), compulsory and a key instrument in performance management by government. We understand that the Ministry of Health and Long-Term Care is seeking to develop methods of performance assessment that emphasise accountability. We have recommended that HRRC continues with its 'soft'

approach, which is what it is designed to do, although individual members of HRRC staff use their considerable expertise to act as advisers on measures proposed by government for accountability.

- v. We believe that Directors and senior managers would find it helpful if HRRC were to produce reports that brings information together on an institutional basis so that they can see where their institution provides models of good practice for the rest of Ontario, and where within their institution there is greatest scope for improvement. It would seem sensible to include information on waiting times in such institutional reports and to develop a standard staff survey to give additional information in *Hospital Reports* and provide a basis for benchmarking for service providers.

Risk adjustment, stratification and choice of benchmark

- vi. The methods of the *Hospital Reports* appear to satisfy recent guidelines for publicly reported outcomes. We have recommended: greater consistency in presentation in relation to benchmarks; additional guidance on what a benchmarking exercise is supposed to represent; and more extensive use of external benchmarks based on judgement (as in the Finance quadrant).

Dealing with both large & small organisations

- vii. There is wide variation in the size of facilities, particularly in CCC and Rehab. We have recommended that a more appropriate way be developed of dealing with small numbers than current practice, which is to omit these from *Hospital Reports*; the use of funnel plots; and ways of handling 'over-dispersion', which occurs, e.g., when a substantial number of hospitals lie outside the funnel limits, indicating that other unmeasured factors other than chance are influencing variability.

Banding of indicators ('performance classification')

- viii. Different *Hospital Reports* and different quadrants within the same *Hospital Reports* use different criteria for banding of an indicator into 'red', 'yellow', 'green. We have recommended a more consistent approach be applied and how banding could reflect the type of indicator.

Indicator aggregation and presentation

- ix. We do not recommend the English system of aggregation to a single summary rating or score for an entire organisation: this is quite inappropriate given the complexity of organisations and the aims of HRRC. Improving consistency in the criteria currently used for banding would improve the use of this information to summarise (or undertake detailed analyses of) performance. It would also be helpful to have clearly specified and consistent criteria, across the different *Hospital Reports*, for identifying outstanding performance. It would also help users of *Hospital Reports* if they followed a common structure (and the model of Rehabilitation, CCC, and ED); had more consistency in the key messages, greater

standardisation in their format, and offered more guidance on the use of indicators for the target audience (as given in the Mental Health Hospital Report).

- x. HRRC reflects common practice, which is to assess acute care by using the indicators that can be developed for a few services given the routinely available data, but there is extraordinary variation in different types of quality of acute care within the same institution. Hence those few indicators tell us about those services only and not the heterogeneous mix of services that make up acute care. We see the key step in developing the next generation of hospital reporting to be producing information that disaggregates acute care into different specific services (e.g. specialties). This raises many problems as disaggregation will produce smaller numbers with more uncertainty and there is a dearth of data on outcomes following hospital treatment other than readmission and death. An interesting development in England has use (by a private insurer) of a standard questionnaire to monitor changes in health status after adult elective surgery

E-scorecard

- xi. The E-scorecard is an exciting development and clearly offers great potential for engaging stakeholders in the available information and providing customised outputs. This could be developed to include the developments we have recommended in *Hospital Reports* of an attractive high-level executive summary for each organisation; use of funnel plots; consistency in banding; and replacing confidence intervals by tolerance intervals around the target, and development of attractive interactive interfaces to increase the use and visibility of the information and as a valuable research project.

Future developments

- xii. Each *Hospital Report* includes an impressive programme of planned developments and we strongly support the proposed development of valid benchmarks and a capacity to drill down from headline indicators. We have also argued for greater consistency in banding and aggregation. The three other kinds of developments we have recommended, were not identified by the different *Hospital Reports*: considering developing of *Hospital Reports* from the point of view of the target audience; the development of attractive interactive interfaces to increase the use and visibility of the information; including information on waiting times; and developing a standard staff survey.

Research priorities

- xiii. We made three recommendations for research: into the development of attractive interactive interfaces; into how the information from *Hospital Reports* is used as a means of enabling their development to have a bigger impact in providers; and exploiting the opportunities offered by the database that has been generated by producing *Hospital Reports* by mining longitudinal data to examine trends over time, and examine optimal size for different services

Report to the Ontario Hospital Report Research Collaborative

Recommendations

Lessons from evaluation of Hospital Reports in Scotland and England for Ontario

- i. We do not know how those who are responsible for delivering health services use the information provided by the different *Hospital Reports*. *We recommend HRRC seek feedback on Hospital Reports from members of boards of directors and senior managers.* We understand that MOHLTC has asked HRRC to do a survey of hospital CEOs. *We recommend that in seeking feedback and in designing this survey HRRC consider the following six criteria* which were developed from a comparison of the impacts of reporting hospital performance in Scotland and England (paragraphs 17 – 27):
 - a) Information ought to be seen as credible and timely (because the data are reliable and not out of date),
 - b) Within organisations there ought to be widespread awareness of the results (they should be easily understood, publicised and widely disseminated);
 - c) The levels of aggregation of indicators ought to relate to responsibility for improving performance;
 - d) There ought to be incentives to act on the information presented and clear accountability to improve poor performance;
 - e) Information ought to be presented to enable benchmarking and learning from the best;
 - f) The information ought to cover what is important (and not just what can be easily measured).

'Hard' and 'soft' approaches in performance assessment

- ii. The 'hard' approach to performance assessment is focussed on accountability and a requirement to meet minimum standards: it is compulsory, about quality assurance, and hence inevitably focused on sanctions for failure. This creates an antagonistic atmosphere between those who assess and those who deliver performance, and hence also inevitably results in gaming. The 'soft' approach aims to enable those who want to improve to learn from the best: it is about quality improvement, focused on success, and hence inevitably voluntary. *We recommend that HRRC continues to practice the 'soft' approach and responsibility in providing information for accountability be taken on by an agency that is part of the government and staff of HRRC be available on an individual basis as expert advisers* (paragraphs 28-29)

Risk adjustment, stratification and choice of benchmark

- iii. The document Hospital Report 2003-2005: First Principles already outlines appropriate overall guidance on a number of topics. *We recommend that additional principles need to be agreed regarding methodology (possibly in a different section) that can be referred to in any situation* (paragraphs 34 –37).
- iv. *We recommend introducing greater consistency in presentation in relation to benchmarks: possibly as observed (O) and expected (E) outcomes under the specific benchmark.* Alternative benchmarks then only influence the expected outcomes (paragraph 39).
- v. Principle 5 in Hospital Report 2003-2005: First Principles establishes an approach to benchmarking. *We recommend additional guidance on what a benchmarking exercise is supposed to represent along the lines of “to compare observed measures with what might be expected were the institution to be performing adequately, where possible taking into account factors beyond the institution’s control”.* Careful consideration clearly needs to be given to the choice of the term used to describe benchmark performance, e.g. “performing adequately / reasonably/ acceptably /” (paragraph 42).
- vi. An Ontario average as a benchmark is generally unsatisfactory and is correctly identified as a last resort. *We recommend the approach developed in the Finance quadrant of extensive use of external benchmarks based on judgement.* However, even with an externally-set benchmark, there may still be a need for statistical methods to deal with chance variability (paragraph 43).

Dealing with both large & small organisations

- vii. The practice of omitting information due to small numbers could encourage gaming such as low survey responses and could conceal disturbing findings. *We recommend that more adequate ways of dealing with small numbers be developed* (paragraph 46).
- viii. Funnel plots essentially plot the indicator against a measure of precision: for proportions this is the denominator, and for O/E this is the expected E. ‘Control limits’ for chance variability can be superimposed. *We recommend the use of funnel plots to provide a natural way of allowing for size of an organisation, and also revealing any association of outcome with size* (paragraph 47)..

Banding of indicators (‘performance classification’)

- ix. *We recommend the use of a general principle for identification of outliers, and hence the banding of an indicator into ‘red’, ‘yellow’, ‘green’.* This could be along the lines of ‘an institution may be identified as an outlier, either high or low, when an indicator shows both statistical and practical significant deviation from a benchmark’ (paragraph 51).

- x. There are currently substantial inconsistencies – both across sectors and quadrants – in the way in which institutions are identified as outliers. This leads to strong variations in numbers being identified as ‘high’ and ‘low’. So a ‘red’ or ‘green’ on an indicator may have different interpretations in different circumstances. This is particularly unfortunate when the E-scorecard uses simple sums of ‘reds’ and ‘greens’ to summarise overall and sector-specific performance. *We recommend a more consistent approach being applied in the different quadrants and sectors in a way that would then allow more consistency in aggregating indicator bands into higher-level categories* (paragraphs 51 and 63).

Aggregation systems: scoring and rules

- xi. As guidance in creating rules for a simple high-level summary, an additional general principle for summarisation may be needed. The current procedures for identifying outstanding performance appears reasonable, but would benefit from a broad description: such as to follow the broad idea of Excellent in many areas, good at nearly all, poor in none or very few. *We recommend such broad descriptions be developed* (paragraph 65).

Variations between Different Hospital Reports

- xii. There are three models of presenting information by the different *Hospital Reports*: Acute Hospital; Rehabilitation, CCC, and ED; and Mental Health. *We recommend that all Hospital Reports follow a common structure and the model of Rehabilitation, CCC, and ED* (paragraph 69).
- xiii. The different *Hospital Reports* emphasise different key messages, have different formats and guidance on use of indicators. *We recommend that all Hospital Reports seek more consistency in the key messages, greater standardisation in their format and offer more guidance on the use of indicators for the target audience (as given in the Mental Health Hospital Report)* (paragraph 71).
- xiv. We would expect that the target audience (of boards of directors and senior managers) would like to see a summary of how their institution performs across all the services they supply. *We recommend that HRRC consider a development that brings together information on different services provided by the same institution in institution-based Hospital Reports* (paragraph 72).

E-scorecard

- xv. We have suggested funnel plots are an attractive way of comparing providers, and these might be considered when generating comparative hospital reports, and in addition to the frequency distribution options for specific indicators and *we recommend the development*

of attractive interactive interfaces to increase the use and visibility of the information and a valuable research project (paragraph 78).

- xvi. A consistent but flexible approach is needed to set tolerances around any benchmark: these tolerances may be constant or reflect size. *We recommend consideration of replacing confidence intervals by tolerance intervals around the target (paragraph 79).*

Future developments

- xvii. Proposed developments reflect the way the work on *Hospital Reports* is organised with considerable autonomy and independence across the different teams. *We recommend HRRC explore another kind of development of Hospital Reports from the point of view of the target audience.* This would form a natural part of a programme to introduce greater consistency in methods and presentation but a focus on helping the target audience use what is produced is also likely to suggest new lines of development: pulling together information from the different *Hospital Reports*, deciding on high level indicators and means of drilling down (paragraph 82).
- xviii. Directors and senior managers need to look at performance in the round and the *Hospital Reports* are the obvious place to do this. *We recommend that HRRC develop Hospital Reports to include information on waiting times (paragraph 83).*
- xix. We understand that in Ontario each organisation may have its own staff survey but that it only allows for examination of changes over time and does not offer a basis for benchmarking. *We recommend that HRRC develop a standard staff survey to offer a basis for benchmarking (paragraph 84).*

Research priorities

- xx. Knowing how information can help local action would give guidance on the development of that information. There are complex issues over interpretation and meanings of differences reported between organisations as identifying a difference as statistically significant is only a start, albeit an important start: we need to understand the reasons for those differences. *We recommend that HRRC develop research into how the information from Hospital Reports is used as a means of enabling their development to have a bigger impact in providers (paragraph 86).*
- xxi. HRRC has built up a considerable data archive. This offers opportunities for analyses of longitudinal data to examine changes over time, develop graphical displays of indicators of direction of travel, analyses of between-hospital variability in trends, and generate hypotheses of drivers of change, which could be explored by qualitative research. The database also offers opportunities for examination of optimal size of different services. *We recommend that HRRC develop research by data mining to examine changes over time*

and optimal size for different services to generate hypotheses to be explored by qualitative analysis (paragraph 87).

- xxii. Analysis of medical practice from small area variations in rates have been used to identify 'supply sensitive' services in the US and a category of high-variation admissions, for which the only plausible explanation was medical discretion in the UK. This research offers a way of raising questions over the appropriateness of utilisation of medical services. *We recommend that HRRRC consider using small area variations as a way of making progress in analysis of acute services (paragraph 88).*

Biographies

David Spiegelhalter

David Spiegelhalter is a statistician in the MRC Biostatistics Unit at Cambridge University specialising in Bayesian methods. He led the statistical team that provided convincing evidence of excessive mortality in paediatric cardiac surgery at the Bristol Royal Infirmary – the landmark case which contributed to the end of self regulation by the medical profession in England, the requirement for the NHS to implement the systems and processes of clinical governance, and the creation of the Commission for Health Improvement (CHI) to review that implementation in England and Wales. David has also developed means of risk adjustment for surgical mortality, was an expert adviser to CHI in the development of performance assessment, and consultant to CHI. He is currently an expert adviser to CHI's successor, the Healthcare Commission, in the development of methods of screening for targeted inspections, surveillance, and monitoring performance against plans.

Gwyn Bevan

Gwyn Bevan is Professor of Management Science at the London School of Economics and Political Science. He was seconded for three years to CHI, where he was Director of the Office for Information on Health Care Performance and had lead responsibility for performance assessment, and, in particular star ratings; national surveys of staff and patients; developing national clinical audits; and undertaking analyses for CHI's reviews, investigations, and national studies. Since leaving CHI he has examined the strengths and limitations of the English system of star ratings and CHI's process of reviewing clinical governance in the NHS.