

FULL REPORT

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

Contents

Summary	4
Hospital Reports in Ontario, Scotland and England	4
Risk adjustment, stratification and choice of benchmark	5
Dealing with both large & small organisations	5
Banding of indicators ('performance classification')	5
Indicator aggregation and presentation	5
E-scorecard	6
Future developments	6
Research priorities	6
Recommendations	7
Lessons from evaluation of Hospital Reports in Scotland and England for Ontario	7
'Hard' and 'soft' approaches in performance assessment	7
Risk adjustment, stratification and choice of benchmark	8
Dealing with both large & small organisations	8
Banding of indicators ('performance classification')	8
Aggregation systems: scoring and rules	9
Variations between Different Hospital Reports	9
E-scorecard	9
Future developments	10
Research priorities	10
1. Introduction	12
2. Hospital Reports in Ontario, Scotland and England	14
Ontario	14
Scotland and England	19
Lessons from evaluation of Hospital Reports in Scotland and England for Ontario	25
'Hard' and 'soft' approaches in performance assessment	26
Presentation for the target audience	27
3. Risk adjustment, stratification and choice of benchmark	28
Introduction	28
Risk-adjustment and stratification	29
Benchmarks	30
4. Dealing with both large & small organisations	30
5. Banding of indicators ('performance classification')	32

6. Indicator aggregation and presentation	33
Aggregation systems: scoring and rules	34
Variations between Different <i>Hospital Reports</i>	35
Disaggregation of acute services	37
7. E-scorecard	37
8. Future developments	38
9. Research priorities	40
10. Conclusions	41
11. References	43
Appendix 1: Schedule of Visit by Bevan and Spiegelhalter	46
Appendix 2: List of CRAG's Clinical Outcome Indicators	48
Appendix 3: Lists of Indicators in Key Targets and the Balanced Scorecard by Type of Organization for Star Ratings Published in 2003	51
Acute	51
Mental Health	53
Mental Health	54
Ambulance Trusts	55
Primary Care Trusts (PCTs)	56
Appendix 4: Comments on specific sectors and quadrants	58
Acute Care	58
Rehabilitation	59
Complex Continuing Care	59
Emergency Department	59
Mental Health	59
Appendix 5: Areas for R &D identified by Report s	60

Report to the Ontario Hospital Report Research Collaborative

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. HRRRC 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).*

Report to the Ontario Hospital Report Research Collaborative

1. Introduction

1. We were asked to review the Hospital Reports with the aim of providing the Hospital Report Research Collaboration (HRRC) with the following terms of reference:
 - i. Feedback and guidance on the statistical methods used in the Hospital Reports;
 - ii. An assessment of the documentation and presentation of methods in the Hospital Reports; and
 - iii. Key methodological research priorities
2. Our experience and expertise is as follows:
 - i. 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.
 - ii. 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.
3. It is helpful here to make some preliminary observations on what we believe we can and cannot usefully offer HRRC and what we aim to avoid.
4. The atmosphere of reporting information in Ontario appeared to us to be very different from that in England. In Ontario the emphasis has been on identifying and celebrating success (as in Scotland, where this approach is described as 'naming and faming'). In the English NHS, the response to scandals over quality of care and unacceptably long waiting times was

to introduce the punitive brutal system of performance assessment of 'star ratings'. This assessed performance of NHS organisations annually against centrally-determined targets (dominated by waiting times) and held chief executives to account for the delivery of Ministerial priorities with sanctions and rewards: a trust that was zero-rated was subjected to the public humiliation of being labelled as a 'failing' organisation and its chief executive faced the threat of being sacked; a 'three-star' trust enjoyed rewards of 'earned autonomy'. Although this was a system of public accountability, the sanctions for failure were much clearer than the rewards for success, and hence the system was described as one of 'naming and shaming'. We are not recommending introducing that approach in Ontario, but do have relevant suggestions based on the English experience of star ratings, and also from the approaches being developed by CHI's successor, the Healthcare Commission (2005a)¹.

5. We would like to emphasise at the start that HRRC is a superb example of a research collaboration that has produced important pioneering *Hospital Reports* in assessing health care performance. We have been enormously impressed by the *Hospital Reports* and the supporting documentation of technical summaries in terms of their organisation, scope and high technical competence. We were equally impressed in our visit (Appendix 1 gives our programme) with our meetings with those working on HRRC, who demonstrated expert knowledge of the different sectors and quadrants, commitment to the endeavour, and enthusiasm to improve and develop methods and presentation.
6. Those who work for HRRC have vastly more knowledge and expertise than we do in each sector and quadrant. We also have not had time to master details of all the material in the *Hospital Reports* and technical summaries. In aiming to offer a stocktake of the HRRC we have sought to look across the current sectors and quadrants to developed generalised principles in methods of analysis and from the point of view of users. We have reviewed variations in methods across sectors and quadrants and propose principles that offer a consistent approach. We have tried to answer questions raised during our visit on the purposes and target audience of the *Hospital Reports*. And we have considered future developments that offer a way of restoring Ontario to its position as innovative leader in use of information to improve performance of health care.
7. The rest of our report is organised into 9 sections which are as follows:
 - i. Hospital Reports in Ontario, Scotland and England
 - ii. Risk adjustment, stratification and choice of benchmark
 - iii. Dealing with both large and small organisations
 - iv. Banding of indicators ('performance classification')

¹ These approaches use the same set of core data for surveillance, inspection and performance assessment and are focused on identifying poor performance. HRRC's emphasis on success, however, requires the same statistical method of identifying 'outliers'.

- v. Indicator aggregation and presentation
- vi. e-Scorecard
- vii. Future developments
- viii. Research priorities
- ix. Conclusions

2. Hospital Reports in Ontario, Scotland and England

8. Ontario was one of the successful pioneers in the publication of *Hospital Reports* (so that anyone working on performance assessment in England in the late 1990s was strongly advised to go and look at what was happening in Ontario). There were similar pioneering developments in the early 1990s in the US (Marshall et al, 2003). This section outlines the development of Hospital Reports in Ontario and compares that with: developments in Scotland, the annual reports of the Clinical Resource and Audit Group (CRAG), from 1994 to 2002; and in England, the annual publication of star ratings, from 2001 to 2005.

Ontario

9. We understand that the underlying principles of the *Hospital Reports* were developed in 1995. The first two *Hospital Reports*, produced in 1998 and 1999, were for Acute Care, sponsored by the Ontario Hospital Association, and produced and based on research by the University of Toronto. From 2001, *Hospital Reports* were sponsored by the Government of Ontario and the Ontario Hospital Association and based on research by HRC. Coverage was extended to: Complex Continuing Care (CCC) and Emergency Department (ED), from 2001²; Rehabilitation, from 2003; and Mental Health, from 2004. *Hospital Reports* are produced by HRC (CCC, ED and Mental Health) and the Canadian Institute for Health Information (CIHI) (Acute Care and Rehabilitation).
10. The emphasis was, and still is, on quality improvement and not accountability. The exercise was, and still is, voluntary and not part of government. HRC includes:
- Canadian Institute for Health Information (CIHI)
 - Centre for Addiction and Mental Health
 - Department of Health Policy, Management and Evaluation, Faculty of Medicine, University of Toronto
 - Department of Rehabilitation Sciences, Faculty of Medicine, University of Toronto
 - Faculty of Nursing, University of Toronto
 - Inner City Health Research Unit, St. Michael's Hospital
 - Institute for Clinical Evaluative Sciences (ICES)
 - McMaster University

² There were also Reports in 2001 on Mental Health, Nursing and Population Health.

- Toronto Rehabilitation Institute
- University Health Network Research Institute
- University of Western Ontario
- University of North Carolina at Chapel Hill
- University of Waterloo

11. The Research Collaborative initially served as both the intellectual engine *and* the production house for the Hospital Reports. Over time CIHI has taken over the production of mature reports and the Research Collaborative focuses on refining existing measures and expanding into new sectors. We strongly support this shift, as it is appropriate for researchers to concentrate on research and development.

12. The objectives of *Hospital Reports* are 'to facilitate local quality improvement programs and to support hospitals' accountability to the communities they serve'. The primary audiences are 'boards of directors and senior managers'; and it is intended that the results be 'shared broadly among hospital staff, patients, families and the public at large'.

13. *Hospital Reports* are organised in the form of a balanced scorecard across four quadrants, and from 2005, a Women's Health Perspective. The information used for each quadrant naturally varies by sector. Performance by hospital was colour coded to show whether this was assessed as being 'above average', 'average' or 'below average'. The information in the Women's Health Perspective was typically reported for the Province rather than at the level of the individual hospital, and examined differences in sex in Patient Satisfaction and for some elements of CUO, and also performance on interventions for women only (e.g., in Acute Care, analyses of labour and delivery, and hysterectomy gynaecological procedures).

14. Table 1 gives details on how information is reported for each quadrant for the different sectors. Information may be organised: by Domain, as for patient satisfaction³; by individual Indicator⁴; and by individual Indicators within Domains⁵.

³ e.g., with Acute Care results from questionnaires are organised into four Domains: Overall Impressions, Communication, Consideration, and Responsiveness.

⁴ e.g., CUO for CCC with improved Activities of Daily Living etc.

⁵ e.g., CUO for Mental Health with results for individual Indicators reported within the Domains of Accessibility, Appropriateness and Outcomes.

Table 1: Sectors and quadrants

	SIC	Patient Satisfaction	CUO	Finance
Acute care	6 Domains	4 Domains	4 Domains & 11 Indicators ⁶	12 Indicators
ED	6 Domains	4 Domains	4 Indicators	4 Indicators
Rehabilitation	12 Indicators ⁷	8 Domains ⁸	3 Domains & 9 Indicators	3 Indicators
CCC	7 Domains	13 Domains ⁹	13 Indicators	8 / 2 Indicators ¹⁰
Mental health	3 Domains & 9 Indicators ¹¹	4 Domains ¹²	3 Domains & 7 Indicators ¹³	7 Indicators

15. The bases for the quadrants in the acute care Report were:

- i. *System Integration and Change (SIC)* covered six areas¹⁴, using data supplied by each hospital completing a self-assessment questionnaire that sought answers to specific behaviours¹⁵. Weighting of questionnaire items was by panels using analogue scales. Performance was assessed as being 'above average', 'average' or 'below average' using two hospital peer groups, teaching / community and small. Great care was taken over checking the quality of data and data entry. This will be improved by the introduction of completion on line, which will also mean that completion of the survey will become less burdensome.

⁶ Three indicators reported by hospital and eight new indicators for the Province.

⁷ There are three Domains for RCGs, Stroke, & Orthopaedic.

⁸ This is described as Client Perspectives

⁹ This is described as Patient & Family Satisfaction with Domains for Patient & Family.

¹⁰ Eight Indicators for free-standing CCCs & two for Acute Care Hospitals

¹¹ With one Indicator under development

¹² This is described as Patient Perception of Care. The Survey to supply these data is under development, only one indicator was reported (discharged against medical advice)

¹³ With two Indicators under development

¹⁴ use of clinical information technology; use of data for decision-making; use of standardized protocols; community involvement and coordination of care; management and support of human resources; and healthy work environment

¹⁵ In 2005, surveys were sent to 123 acute hospitals and were returned by 108 hospitals (a response rate of 88%).

- ii. *Patient Satisfaction* covered four areas¹⁶ using results from a survey by National Research Corporation and Picker (using questions similar to those used in the US) of patients who had an acute inpatient stay, during two different time periods¹⁷. The rule for inclusion varied across sectors: this was, e.g., a hundred completed surveys for acute care but lower for CCC. There are opportunities for bias from selection of which responses to return, and incentives to return a small number (as hospitals have to pay for the survey and a small sample size will be associated with wide confidence intervals). Answers were adjusted for differences in age and sex. The reported statistics are weighted averages by hospital. There are problems over correlations between measures, and hence of aggregation.
- iii. *Clinical Utilization and Outcomes* (CUO) covered three Case Mix Groups¹⁸ and used 11 indicators of adverse events, readmissions and appropriateness based on the results of a comprehensive literature review and the advice of expert panels.
- iv. *Financial Performance and Condition* covered five broad categories of 12 indicators¹⁹ with most indicators derived from data in an electronic database²⁰. Although these data ought to be reliable, they are subjected to further checks. There is a time lag of a year and a half between the collection of the data and the publication of the report. Expected costs were estimated by a regression model (with questions over whether there are diseconomies of scale and the grouping of small hospitals). There were no obvious bands of performance. For the two new indicators introduced in the *Hospital Report* for 2005 benchmarks were derived from a survey of chief financial officers. The objective was to identify excellence, with high performance being defined for 2003-04 as achieving above average performance on 9 of the 12 indicators.

16. Tables 2, 3, 4 and 5, which were prepared by HRRC, give the data structure for each quadrant by sector and show that the quadrants in the different sectors follow the model of acute care²¹.

¹⁶ Overall Impressions, Communication, Consideration, and Responsiveness

¹⁷ during 2003–2004, and several months of 2004–2005

¹⁸ Medical, Surgical or Major Surgical

¹⁹ one of financial viability (expected long-term financial health); three of efficiency (ratios of hospital outputs to costs); two of liquidity (management of current assets and liabilities); one of capital (maintenance of long-term assets); and five of human resources (allocations to patient care and other activities)

²⁰ that contains the internally generated year-end general ledger balances for each hospital in the province

²¹ For Patient Satisfaction in CCC the surveys are based on interviews with residents, which causes problems from small numbers.

Table 2: SIC Data Structure

Sector	Indicator Format	Data Source	Unit of Obs	Sample / Population
Acute	Scale (0-10)	Survey Questions	Hospital	Population
ED	Scale (0-10)	Survey Questions	Hospital	Population
CCC	Scale (0-10)	Survey Questions	Hospital	Population
Rehab	Scale	Survey Questions	Hospital	Population
Mental Health	Scale	Survey Questions/ DAD	Hospital	Population

Table 3: Patient Satisfaction Data Structure

Sector	Indicator Format	Data Source	Unit Obs	Sample / Population
Acute	Scale from (0-100)	Survey questions (NRC-Picker)	Patient	Sample
ED	Scale from (0-100)	Survey questions (NRC-Picker)	Patient	Sample
CCC	Scale from (0-100)	Survey questions (NRC-Picker)	Patient	Sample
Rehab	Scale from	Survey questions	Patients	Sample
Mental Health	Rate	DAD	Patients	Sample

Table 4: CUO Data Structure

Sector	Indicator Format	Data Source	Unit Obs	Sample/ Population
Acute	Rate (DAD)	Administrative	Patient	Population
ED	Rate (NACRS/DAD)	Administrative	Patient	Population
CCC	Rate (CCRS)	Administrative	Patient	Population
Rehab	Scale Or Average stays	Administrative (NRS)	Patient	Population
Mental Health	Rate	Administrative (DAD)	Patient	Population

Table 5: Finance Data Structure

Sector	Indicator Format	Data Source	Unit of Obs	Sample /Population
Acute	Continuous (Ratio)	OHS (routinely collected administrative data)	Hospital	Population
ED	Continuous (Ratio)	OHS (routinely collected administrative data)	Hospital	Population
CCC	Continuous (Ratio)	OHS (routinely collected administrative data)	Hospital	Population
Rehab	Continuous (Ratio)	OHS (routinely collected administrative data)	Hospital	Population
Mental Health	Continuous (Ratio)	OHS (routinely collected administrative data)	Hospital	Population

Scotland and England

17. As mentioned above, HRR looks to us to have strong similarities with the approach taken in Scotland in the development of performance assessment and be very different from that developed in England. Carter et al. (1995: p. 49) pointed out that health services are

characterised by 'heterogeneity, complexity and uncertainty, and hence most performance indicators are "'tin openers" rather than "dials" they do not give answers but prompt investigation and inquiry, and by themselves provide an incomplete and inaccurate picture'. This message is consistently emphasised in the different *Hospital Reports* produced by HRRC, that: the commonly accepted statistical techniques used do not entirely eliminate the impact of uncontrollable factors on indicator results; indicators ought to be viewed as screening tests that can identify potential opportunities for quality improvement; there is a need for hospitals to "drill down" using their own data to better understand the factors underlying their results; and the e-Scorecard offers a means of doing so. CRAG published 'tin openers' (clinical indicators) and like HRRC emphasised their limitations as a guide to action. In England a very different system of annual performance assessment was introduced in 2001: the publication of a single summary score of star ratings (from zero to three stars), which were dominated by 'dials' (waiting times) linked with sanctions and rewards.

18. CRAG published clinical outcome indicators between 1993 and 2001 (Appendix 2 gives the list of the indicators used over this time). The impact and usefulness of these *Hospital Reports* were evaluated by independent academics based in England (Mannion and Goddard 2001), who explored their impact on NHS trusts in Scotland; and by a CRAG-funded Clinical Indicators Support Team (CRAG 2002: pp. 223–228), who investigated the requirements of health boards and trusts for clinical performance information, whether the indicators met their needs, and how and why the indicators have been used. The main conclusion of the evaluation by Mannion and Goddard (2001: p. 260) was that in Scottish trusts these indicators 'had a low profile ... and were rarely cited as informing internal quality improvement or used externally to identify best practice'. The Clinical Indicators Support Team (CRAG 2002) came to similarly depressing conclusions.
19. Star ratings were published from 2001 to 2005²², and from 2003 covered all five types of NHS trusts: acute, specialist, mental health, ambulance, and primary care. Although the system of star ratings was portrayed as giving the public and patients a rounded assessment of NHS performance, its actual objective was to use publication to put pressure on those within the NHS to reduce long waiting times (Bevan, and Hood 2006a and b). This punitive system was coupled with the Modernisation Agency (Secretary of State for Health, 1997 and 2000), which offered support to NHS so that they could learn from best practice to help organizations and which identified '10 High Impact Changes' (Modernisation Agency, 2004).

²² The Department of Health developed and published star ratings in 2001 and 2002 (Department of Health 2001, 2002a). From 2003, responsibility for the development and publication of star ratings passed to independent regulators, CHI for 2003 (Commission for Health Improvement 2003a and b), and CHI's successor, the Healthcare Commission for 2004 and 2005 (Healthcare Commission 2004, 2005b).

20. The model used for star ratings of acute trusts in 2001, was adapted for the other types of organizations, and for the first three years used three different kinds of indicators:

- Assessments from CHI's Clinical Governance Reviews (CGRs) on the seven technical components of clinical governance (see Figure 1), to avoid conflicting assessments of performance²³.
- Nine 'key targets': six were waiting times, and the other three were achieving a financial balance, hospital cleanliness, and improving the working lives of staff (Appendix 3 gives details of key targets for 2003).
- About forty targets organised in three focus areas of a (so called) 'balanced scorecard' that reflected Ministerial priorities as promulgated in a much larger number of targets in the Priorities and Planning Framework (PPF) (Department of Health 2002b); and satisfied technical criteria of being applicable nationally, measurable, capable of being captured by indicators, and stable over time (Appendix 3 gives details of the 'balanced scorecard' for 2003).

Figure 1: Technical Components of clinical governance

	Component
Resources and processes processes for quality improvement	Patient and public involvement
	Clinical audit
	Risk management
	Clinical effectiveness programmes
staff focus	Staffing and staff management
	Education, training and continuing personal and professional development
Use of information	Use of information to support clinical governance and health care delivery

²³ The Government defined Clinical Governance as 'a framework through which NHS organizations are accountable for continuously improving the quality of their services and safeguarding high standards of care by creating an environment in which excellence in clinical care will flourish'. CHI's tasks included undertaking Reviews of clinical governance and producing reports of those reviews. These reviews provided information on the implementation of seven technical components of clinical governance, which were scored by CHI: I (little or no progress), II (implementation in part), III (substantial progress) and IV (excellent). It could, and did, happen that trusts performed well on the 'key targets' and 'balanced scorecard', but made limited progress in implementing clinical governance, and vice versa.

21. Scores from CHI's CGRs and performance against 'key targets' were given priority in the star rating system and hence by Chief Executives in the NHS. So what was included in ratings defined what mattered, and hence by implication, what was excluded did not matter. Ratings were designed with the intention of achieving Ministerial priorities (in particular reducing long waiting times) subject to satisfying important constraints (e.g. cleanliness and avoiding financial deficits). The total number of targets was limited to 50 (because it was unacceptable to hold NHS chief executives to account for a vast number). The role of the 'balanced scorecard' was symbolic: to justify the claim that star ratings offered a rounded assessment of NHS performance, and hence that zero-rating meant that an organisation was 'failing' and was justly 'named and shamed'. The implications were that the 'balanced scorecard' was required to include a heterogeneous mix of targets, but its actual composition was of lesser importance (Bevan, 2006). Furthermore, the scoring system of star ratings worked in a sequence, in which 'key targets' provided the first hurdle. So being zero-rated was a consequence of failure against the small set of 'key targets' only, and that failure could not be redeemed by outstanding performance in the 'balanced scorecard'.
22. There is evidence of two kinds of impacts of star ratings: reported improvements against 'key targets' and of gaming²⁴ (Bevan, 2006; Bevan and Hood 2006a and 2006b). We give two examples that illustrate both impacts.
- i. A key target for ambulance trusts was that *75% of immediately life-threatening emergencies (category A calls) being met within 8 minutes*. This target had existed since 1996. Figure 2 shows that after this became a key target for ambulance trust star ratings in 2001/2, reported performance jumped dramatically. Although a number of Trusts failed to achieve this target, there was a notable change, in that for the pre-star-rating year ending in March 2000, some Trusts only managed 40 per cent, but three years later, the worst achieved nearly 70 per cent. A study by the Commission for Health Improvement (2003c), however, found evidence that in a third of ambulance trusts, response times had been 'corrected' to be reported to be less than eight minutes. This pattern is shown in Figure 3, which gives an example of the expected pattern where there has been no 'correction', and of a correction that produced the curious 'spike' at 8 minutes – with the strong implication that times between 8 and 9 minutes have been reclassified to be less than 8 minutes.

²⁴ which is known to have been endemic when targets were used in centrally planned economies (Nove 1958, Kornai 1994) and in the public sector (Smith 1995)

Figure 2: Percentages of category A calls met within 8 minutes

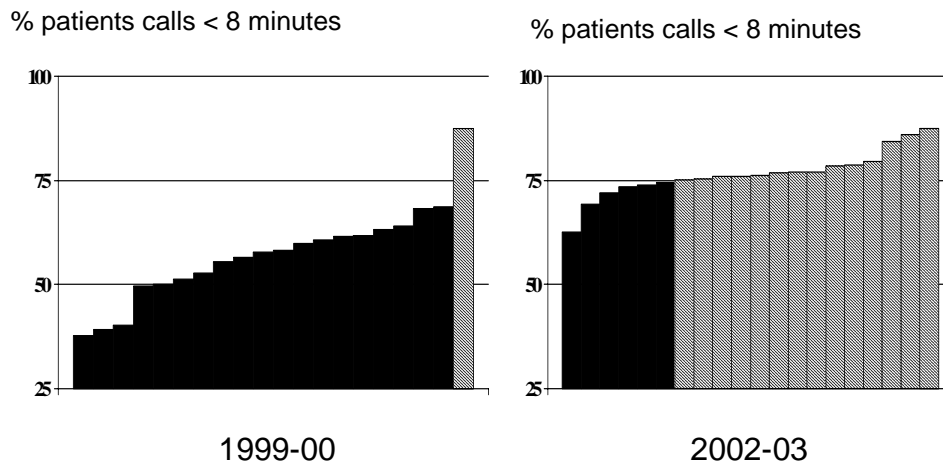
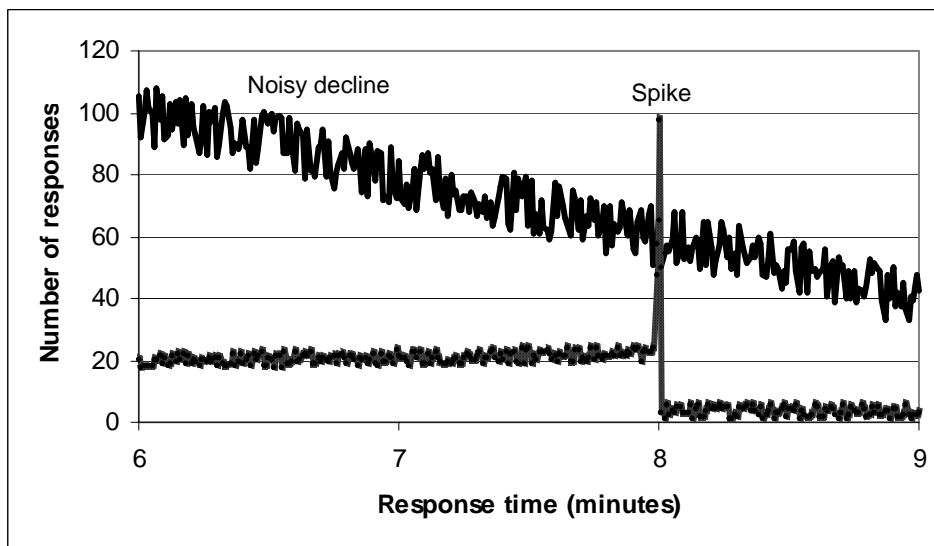
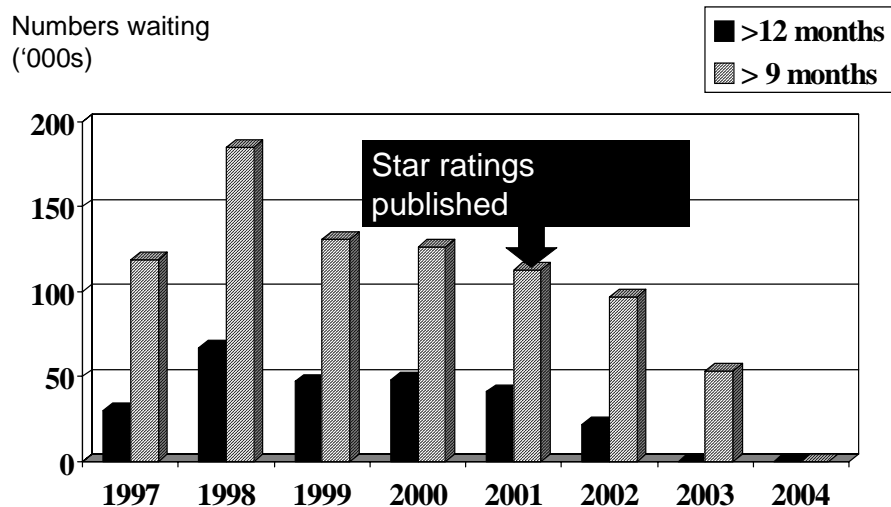


Figure 3: Frequency distributions of ambulance response times



- ii. A key target for acute trusts was the *maximum waiting time for first elective admission: 18 months by the end of March 2001, 15 months by 2002, 12 months by 2003, and 9 months by 2004*. Figure 4 shows the dramatic impact of these targets. But Auditors found 12 trusts had fiddled the statistics (National Audit Office 2001, Audit Commission 2003); and other gaming practices included removing patients from inpatient waiting lists once they had been provided with a future date for an appointment, giving patients immediate appointments that they were not able to attend (and who were then classed as refusing treatment), and inappropriately suspending having treatment (Commission for Health Improvement 2004). There is also an implication that the system encouraged trusts to spend their way out of trouble in waiting times even though this resulted in financial deficits: about a third of acute trusts 'significantly underachieved' their key financial target in the final year of star ratings (Bevan, 2006) ²⁵.

Figure 4: Numbers waiting for elective admission in England



²⁵ The technical basis of scoring in star ratings allowed acute trusts to make trade-offs to avoid being zero rated by overspending to meet the six waiting time key targets. The problem of financial deficits became acute in 2006: it was seen to be the primary cause of the resignation of the Chief Executive of the NHS in England, and is the subject of an Inquiry by the Parliamentary Select Committee on Health.

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

23. There are striking differences between the coverage of *Hospital Reports* in Scotland, England and Ontario:
- i. Ontario *Hospital Reports* have a well-developed assessment of financial performance. This was omitted from CRAG's *Hospital Reports* and inadequately covered by star ratings (with only one 'key target' for financial balance).
 - ii. Star ratings were dominated by waiting time targets, which are omitted from CRAG's Reports and HRRC's *Hospital Reports*.
 - iii. Star ratings and CRAG's *Hospital Reports* included indicators for public health and primary care, for which there appear as yet to be no *Hospital Reports*.
 - iv. Star ratings included assessments of implementation of clinical governance²⁶ (Scally and Donaldson 1998) which like SIC emphasise the systems and processes to support quality improvement, good use of information and guidelines, and the importance of looking after and developing staff. SIC emphasises coordination across sectors and the community, which does not explicitly feature in the elements of clinical governance; and clinical governance includes patient and public involvement, clinical audit, and risk management, which do not explicitly feature in the elements of SIC.
24. Evaluations of the Scottish exercise and comparisons with star ratings suggested a check list of six criteria for systems of performance assessment to have an impact:
- i. Information ought to be seen as credible and timely (because the data are reliable and not out of date),
 - ii. Within organisations there ought to be widespread awareness of the results (they should be easily understood, publicised and widely disseminated);
 - iii. The levels of aggregation of indicators ought to relate to responsibility for improving performance;
 - iv. There ought to be incentives to act on the information presented and clear accountability to improve poor performance;
 - v. Information ought to be presented to enable benchmarking and learning from the best;
 - vi. The information ought to cover what is important (and not just what can be easily measured).

²⁶ The government defined clinical governance as 'a framework through which NHS organisations are accountable for continuously improving the quality of their services and safeguarding high standards of care by creating an environment in which excellence in clinical care will flourish' (Department of Health, 1998).

25. CRAG published 'tin openers' (clinical indicators), failed all six criteria and was ineffective. In contrast, star ratings were dominated by 'dials' (waiting times), and with the Modernisation Agency satisfied the first five criteria, and looked to be effective. The weaknesses of star ratings were: gaming, in response to 'key targets'; and the coverage in the 'balanced scorecard', which did not satisfactorily resolve the formidable problems of capturing performance of complex organisations in forty additional indicators (Bevan, 2006).
26. We do not know how those who are responsible for delivering health services use information in the *Hospital Reports* in Ontario against. The impression that we gained during our visit was this is seen as credible because the data are reliable, but there have been reservations about timeliness because of time lags in the collection and analysis of data, and reporting of results. We understand that MOHLTC has asked HRRC to do a survey of hospital CEOs. *We recommend that those designing this survey consider exploring the criteria listed in paragraph 22 and HRRC also seeks feedback on Hospital Reports from members of boards of directors and senior managers.*
27. We discuss here the different emphases on approaches to reporting performance illustrated by Scotland and England, questions of presentation for the target audience, and variations between *Hospital Reports* for the different sectors. We discuss below issues of aggregation, coverage, incentives and accountability.

'Hard' and 'soft' approaches in performance assessment

28. 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. In practice these overlap: the English system of star ratings was supported by the modernisation agency to enable those who were failing to learn from the best; and a commitment to quality improvement does not justify turning a blind eye to scandalously poor quality of care. Nevertheless it is clear that, star ratings exemplified the 'hard' approach, and CRAG and HRRC the 'soft' approach.
29. It is difficult for the same organisation to practice both the 'hard' and 'soft' approaches. Producing information for quality improvement requires an atmosphere in which those who deliver health care are willing to be open about where their performance needs to be improved, but using information for accountability sours that atmosphere. We understand that the MOH is appropriately committed to developing a hard approach that will hold health care providers to account for their performance. This raises a question over the role of HRRC in the development of this hard approach. If HRRC were to move to providing information for the government as part of these new developments in accountability, HRRC

would in effect cease to be able to function properly in producing information for improvement, which is what it is precisely designed to do: it is run on a voluntary basis, led by University, and at arms length from Government. HRRC staff have considerable expertise in assessing performance of health care and individuals can usefully act as advisers on the measures proposed by government for accountability. This would still mean that HRRC is at arms length from Government (which may or may not take this advice). *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.*

Presentation for the target audience

30. Without going to the extreme of the high-powered incentives of star ratings, in which information on performance was directly linked with sanctions and rewards, it is helpful to create an atmosphere that not only celebrates success, but also encourages improvement where that is needed. As Florence Nightingale observed, reports are not self executive (Woodham-Smith, 1970).
31. Information on performance assessment may be used by different audiences. The primary audiences for HRRC are boards of directors and senior managers, but it is also hoped that the results will be shared with hospital staff, patients, families and the public at large (no mention is made of government or the media, or clinical staff). Different audiences do, of course, have different needs. HRRC is not designed to meet the needs of patients and their families, who want timely specific information relevant to treatment of the patient's condition (such as the type of surgical procedure). So we see the target audience being those responsible for governance and management of hospitals, which includes boards of directors, senior managers, and clinicians. We have recommended above having feedback from members of boards of directors and senior managers on Reports. A key issue in England has been providing an independent source of information for non-executive directors and we wonder if this is a group that HRRC might target.
32. We see advantages in seeking to inform the public through the media. We understand that the high performers in the Finance Quadrant made the front pages of their local newspapers. Hibbard et al (2003) report from an experiment on reporting information on hospitals in Wisconsin that only in hospitals where information was published managers did use information to improve services; and in hospitals here information was not published no improvements were observed whether they had been supplied with information or not. And the Clinical Indicators Support Team (CRAG 2002) pointed out that the early CRAG Reports 'had more impact in the media and this put pressure on trusts and health boards' and that without the that coverage in later Reports, this pressure was absent.

33. To create an atmosphere that generates pressure for improvement we see advantages in providing a simple high-level summary to set the agenda. This could be organised to enable a member of the Board to answer three kinds of questions, namely for service X, is our *hospital* providing:
- a) The best quality in Ontario and a model for other to follow?
 - b) High quality care alongside other hospitals in Ontario and we have little to learn by prioritising this specialty?
 - c) Providing care of lower quality such that we ought to prioritise this service for improvement?

3. Risk adjustment, stratification and choice of benchmark

Introduction

34. The next sections largely deal with methodology and involve some technicalities: we give examples of specific issues in the main text and Appendix 4 provides more detailed comments on specific sectors and quadrants.
35. 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.* We shall outline some broad ideas, but of course these would need to be decided by the appropriate bodies.
36. HRCC appears to satisfy the recent AHA Guidelines for publicly reported outcomes (Krumholz et al, 2006) which are summarised as:
- a. Clear and explicit definition of an appropriate patient sample;
 - b. Clinical coherence of model variables;
 - c. Sufficiently high-quality and timely data;
 - d. Designation of an appropriate reference time before which covariates are derived and after which outcomes are measured;
 - e. Use of an appropriate outcome and a standardized period of outcome assessment;
 - f. Application of an analytical approach that takes into account the multi-level organization of data; and
 - g. Disclosure of the methods used to compare outcomes, including disclosure of risk-adjustment methodology in derivation and validation samples.
37. The guideline on multilevel modelling (item f) raises a controversial issue: whether to use a hierarchical model in which hospital is considered as a random effect, leading to estimates of the hospital effect that are “shrunk” towards the overall mean. This is perhaps best

thought of as an adjustment for regression-to-the-mean: essentially, by viewing each hospital as a member of a population, it is reasonable to believe that extreme results relative to other hospitals in one year may be partly due to a run of bad or good luck, and next year the results will tend back towards the overall level. Such techniques should therefore generally produce better predictions of the next year's performance. Currently the HRCC uses these techniques for surveys, using a regression analysis with hospital as random-intercept. These methods are fine if concerned with estimating long-term risk levels, or predicting next year's performance, but there is a risk of losing the 'face-validity' of using simple Observed and Expected outcomes.

Risk-adjustment and stratification

38. Every effort should be made to make some type of adjustment for factors over which the hospital has no control, if only to improve the face-validity of the comparisons. In the clinical context, such factors are measured at the *patient* level and seek to describe the 'case-mix' in terms of age, sex, co-morbidities and so on; in non-clinical measures such as SIC such factors are measured at the *institutional* level and could relate to aggregate description of the patient population, size and type of hospital and so on. At the patient level, 'risk-adjustment' generally uses a statistical regression model to adjust outcomes for case-mix, while 'risk-stratification' delineates separate patient risk-groups within which observed outcomes are compared with benchmarks, and then the comparison aggregated over strata. Which is best depends on circumstances and data available: simple risk-stratification will often be adequate and be more generally applicable. At the institutional level, it will generally only be possible to present an analysis stratified for institutional type: while currently a division into 'teaching', 'small' and 'community' is used, we suggest a review of whether this is an appropriate stratification in all circumstances.
39. *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.*
40. For rare (generally adverse) events, indirect standardisation (say using a logistic regression model) can provide expected outcomes. Age/sex would be minimum risk factors (except when explicitly comparing male/female outcomes). Results can be reported as O/E, or O/E x overall rate, as a 'risk-adjusted' rate as in New York cardiac report cards (http://www.health.state.ny.us/nysdoh/heart/pdf/2001-2003_cabg.pdf), and possibly plotted using funnel plots (see below).
41. Another way of handling very rare events is to use a very different approach of confidential inquiries to provide information for learning from mistakes.

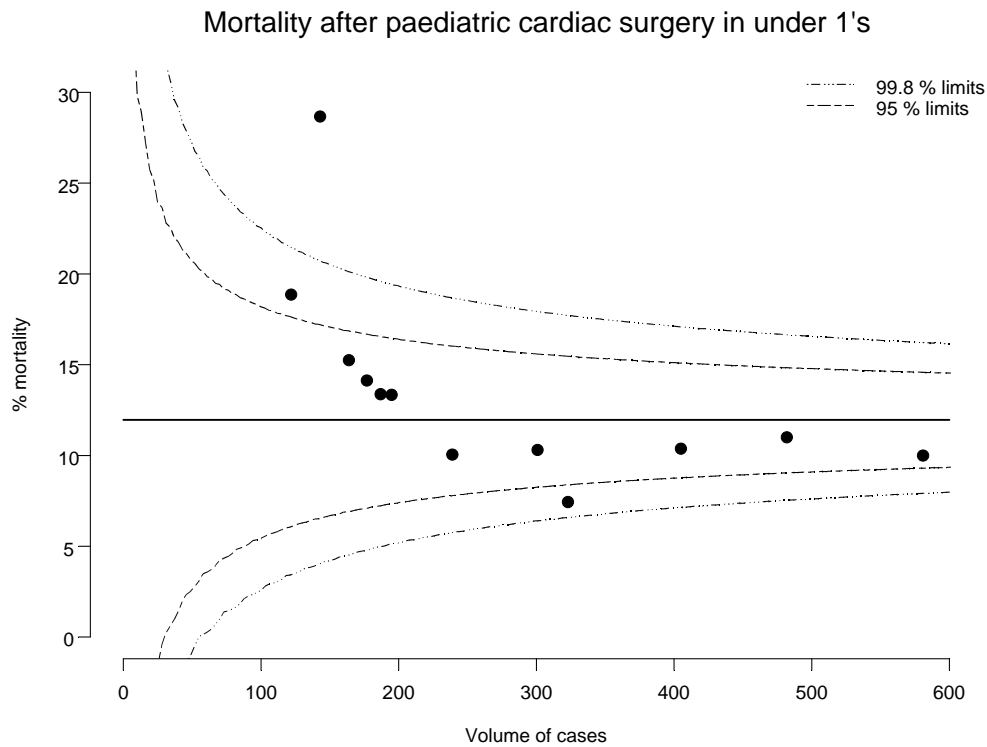
Benchmarks

42. 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 hospital to be performing adequately, where possible taking into account factors beyond the hospital '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 /...."
43. 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 (see below).
44. It is difficult in Complex Continuing Care (CCC) to identify homogenous populations to compare and hence to create benchmarks. In this circumstance it may be appropriate to use the facility's own baseline as a "local benchmark" and hence look for intra-facility change. This is more like classic industrial control charting in looking for changes over time, and is reasonable with, say, good clinical data for at least 4 years. This is an example of the use of longitudinal data discussed further below.

4. Dealing with both large & small organisations

45. There is wide variation in the size of facilities, particularly in CCC and Rehab. This raises the issues of (a) how to deal with both large and small organisations, (b) researching the possible influence of size.
46. 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*.
47. Funnel plots (Spiegelhalter, 2005a) 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. Figure 5 is an example taken from the Bristol Royal Infirmary Inquiry (Spiegelhalter, 2002), which not only clearly shows that Bristol was an outlier, but also the volume effect for the other hospitals. *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*.

Figure 5. Funnel plot of mortality rates of 12 English hospitals carrying out paediatric cardiac surgery 1991-1995



48. 'Over-dispersion' is an important issue (Spiegelhalter, 2005b): this occurs when a substantial number of hospitals lie outside the funnel limits, indicating that other unmeasured factors other than chance are influencing variability. We can think of this as a degree of unavoidable variability that will lead large *hospitals*, even after risk-adjustment, to display some variability. Ignoring this can lead to a bias against large hospitals, since differences might be identified as statistically significant, but which are not practically significant. In for example, CCC, this has led to the "low cut-off" methodology that avoids penalising organisations that may be statistically, but not practically, below a threshold. This can be dealt with in two ways: (a) setting tolerance thresholds around benchmarks which do not depend on sample size (as in Type B indicators described below), or (b) a formal but simple over-dispersion technique. For example, for patient surveys, allowing for over-dispersion would just add the between-hospital variance to the within-hospital variance when calculating the P-value.

49. The issue of size of facility is clearly an important research area, as shown in the Bristol funnel plot shown above. This could be investigated within CCC and other sectors: a relationship with size is both of interest in itself and because it may then be reasonable to include size as a stratifier / predictor factor.

50. For indicators measured at the institutional level, such as in Finance and SIC, there is no natural way of measuring 'expected' variability and so setting control limits. Nevertheless it is still natural to plot outcomes against institutional size in order to illuminate any relationship. Agreed 'tolerances' (see discussion below on 'bandings') can still be shown, even if these are horizontal thresholds that are independent of size.

5. Banding of indicators ('performance classification')

51. *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 hospital may be identified as an outlier, either high or low, when an indicator shows both statistical and practical significant deviation from a benchmark'.*

52. We mentioned above the issue concerning the current substantial inconsistencies – both across sectors & quadrants – in the way in which hospitals are identified as outliers. This leads to strong variations in numbers being identified as 'high' and 'low' (for example SIC in Rehab has a majority of cases being 'high' or 'low!'). Specific issues are discussed in Appendix 4. *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.*

53. Banding indicators is essentially an issue identifying reasonable tolerances around a benchmark. Such tolerances clearly depend on the type of indicator, and one might identify three broad types of indicator (see Table 6).

Table 6: Types of indicators

<i>Type of indicator</i>	<i>Degree of control of organisation on precise outcome</i>	<i>Examples of indicator</i>	<i>Degree of tolerance around benchmark</i>
A	Strong	Employee numbers	Zero
B	Moderate – substantial influence of external factors	X-ray rates, delayed discharges	Size-independent limits based on judgement or empirical distribution
C	Weak – substantial influence of chance	Surgical outcomes	2 or 3 Standard deviations (possibly allowing for over-dispersion)

54. While organisations clearly can influence the overall level of all types of indicators by changes in structure and process, the crucial issue in distinguishing the type of indicator and the consequent tolerances is the extent to which the precise outcome is controllable.
55. On a funnel plot, type B indicators correspond to horizontal thresholds, while type C the standard funnel. If there is substantial over-dispersion it points to the influence of risk factors that have not been taken into account in the benchmarking, and the resulting tolerances form a compromise between horizontal and funnel limits.
56. The Healthcare Commission website <http://heartsurgery.healthcarecommission.org.uk/> shows survival for named hospitals and surgeons. A simple banding is based on whether the observed rate falls outside a 3 SD interval around the expected rate – this can be thought of as a cross-section of a funnel plot.
57. Any indicators with an externally- or judgementally-set benchmark could have a large number of high / low performers.
58. For Type B indicators, thresholds have to be set from judgement or based on distribution of outcomes across *hospitals*. To avoid a system that is purely 'relative', e.g. always identifies the top and bottom 10% of *hospitals* as 'red' or 'green', it may be appropriate to use the distribution of a previous year's results. In this way there is a potential for general improvement.
59. Clearly the 'types' shown in Table 6 are not an absolute categorisation: for example, in the Healthcare Commission it has been found that for certain process measures such as 'numbers of drug users completing rehab programmes' it can be argued that there is a substantial contribution from both chance and the quality of the programme. The Type B indicators are generally of this type, where some tolerance is reasonable but not dependent on size of the institution. Nevertheless we feel that explicit, structured discussion of permitted tolerances is a valuable tool that should bring additional coherency to a currently rather disparate approach.

6. Indicator aggregation and presentation

60. We see the issue of aggregation and disaggregation as a key issue in framing the future developments of HRRC and as offering a means for restoring Ontario to its position as innovative leader in use of information to improve performance of health care. We have emphasised advantages of a simple high-level summary within each *Hospital Report*, but that can, however, only highlight areas for further investigation. It is vital to be able to drill down to extract details for those who need to learn and act. We consider here three different sets of issues:
- i. Aggregation systems: scoring and rules,

- ii. Variations between Different *Hospital Reports* and
- iii. Aggregation and disaggregation.

Aggregation systems: scoring and rules

61. 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.
62. Scoring systems summarise a set of indicators by adding up points, while rules provide a summary by a set of logical criteria (e.g. 9/12 indicators 'above average' and none 'poor'). The primary difference is that scoring systems allow trade-offs, so that very good performance on many indicators can overcome poor performance on some, while rules may require a 'minimum' level across all indicators before labelling as 'high performing'. An additional advantage of scoring-systems is that they can be consistently applied regardless of the number of indicators.
63. We mentioned above the lack of consistency across sectors and quadrants, and 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. This is another reason for our above recommendation for greater consistency in identification of outliers.
64. It seems reasonable to adopt simple additive scoring systems at a low-level when combining correlated items measuring similar constructs: this essentially produces an 'average' score within a particular area, which should be reasonably interpretable. The appropriate role of rule-based systems is for high-level aggregation where judgement is expected to play a part: at this level it would not generally be reasonable to allow good performance in some areas to 'cancel out' real failures in others. This is how they are currently used in the Annual Health Check of the Healthcare Commission.
<http://www.healthcarecommission.org.uk/serviceproviderinformation/annualhealthcheck.cfm>.
65. 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: for example to follow the general idea of "Excellent in many areas, good at nearly all, poor in none or very few". *We recommend such broad descriptions be developed.*
66. Benchmarking against external (rather than statistical) standards may result in many 'red' or 'green' organisations, and will emphasise the need to be able to drill down.

Variations between Different *Hospital Reports*

67. It is acknowledged that the way in which HRRC extended *Hospital Reports* into other sectors was to allow each team considerable autonomy to develop what they deemed best to capture the services of each sector with the data that are routinely available are could be readily collected (through surveys). In addition to the differences in how benchmarks were set there is also variation in the organisation of information in the different *Hospital Reports* and in key messages.
68. The organisation of information for the *Hospital Reports* for Rehabilitation, CCC, and ED had a common structure:
- i. *Scorecard Overview*, which gave answers to three questions:
 - i. What does the scorecard illustrate?
 - ii. How can hospitals use the results?
 - iii. Do the scorecard results relate to key strategic priorities?
 - iv. Are there high performing hospitals?
 - ii. *Background*, which described the coverage of the Report in terms of care provided and numbers of hospitals included out of the total in Ontario;
 - iii. *A Balanced Scorecard*, which gave the definition of each quadrant and the Women's Health Perspective and the colour coding used for each, and that for ED usefully gives the numbers of indicators in each quadrant;
 - iv. *Interpreting Scores*, which explained the definitions of the banding used;
 - v. *Information for each quadrant*: which gave definitions of indicators, results for the Province and in detail by hospital for each Domain or Indicator (classifying each hospital as 'above average', 'average' or 'below average'); and
 - vi. *Women's Health Perspective*, which included analyses of Patient Satisfaction and CUO.
69. The Acute Hospital Report was similar to that for Rehabilitation, CCC, and ED but did not have the section on 'A Balanced Scorecard'²⁷. The Mental Health Report had no colour coding and a different layout, which described in details the indicators being considered as well as giving results on a regional basis only. *We recommend that all Hospital Reports follow a common structure and the model of Rehabilitation, CCC, and ED.*
70. The different *Hospital Reports* emphasise different key messages, have different formats and guidance on use of indicators.

²⁷ And for acute care patient satisfaction precedes CUO and for the others this order is reversed.

- i. *Acute Care*²⁸: 'To ensure optimal use of the scorecard results, board members should identify indicators for which their hospital's performance is lower than average or for which sex differences are significantly different and ensure that sufficient resources are allocated to facilitate quality improvement in these areas'.
- ii. *Emergency Department Care*²⁹: To achieve the objectives of this Report, Boards of Directors, and senior managers should identify meaningful ways to share the results with middle management, decision support and quality improvement staff, front line staff, patients, families, emergency service networks, base hospitals, and their communities.
- iii. *Rehabilitation*³⁰ 'Hospital managers can use these reports to identify other hospitals from which they might seek opportunities to learn'.
- iv. *CCC*³¹ 'Taken as a whole, the balanced scorecard for CCC provides a summary framework for improving hospital efficiency and clinical outcomes. Its purpose is to offer a 'big picture' approach to effective CCC services. This **Report** does not aim to provide advice to patients trying to decide where to go for care.
- v. *Mental Health*³² is the only *Hospital Report* that has a whole section devoted to guidance on how to use the indicators and covers:
 - Contextual and Grouping Variables: explaining what each aims to measure (e.g. capacity or flowthrough)
 - 'Drill Down' Example: of Indicator SIC9, Notification of Hospitalization,
 - Cross-Quadrant Analysis Example: relationships between rate of OHIP follow-up care within 30 days of discharge, and patient participation in discharge planning, and physician availability.
 - Broad Picture Summary: in terms of Appropriateness, Outcomes, Participation and System Management

71. *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 Report).*

72. We would expect that the target audience (of boards of directors and senior managers) would like to see a summary of how their *hospital* performs across all the services they

²⁸ Interpretation of the Results (page 8)

²⁹ How Can Hospitals use the Results? (page 5)

³⁰ How Can Hospitals use the Results? (page 4)

³¹ How Can Hospitals use the Results? (page 5)

³² Facilitating Interpretation (pages 87-94)

supply. *We recommend that HRRC consider a development that brings together information on different services provided by the same hospital in different Hospital Reports.*

Disaggregation of acute services

73. The 2005 Hospital Report for Acute Care has three indicators of Clinical Utilization & Outcomes:
- i. medical readmissions(for acute myocardial infarction (AMI), heart failure, asthma, gastrointestinal bleed and stroke),
 - ii. surgical readmissions (for cholecystectomy, hysterectomy and prostatectomy), and
 - iii. surgical appropriateness (the percentage of cholecystectomies performed “open” versus laparoscopically).
74. 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 (although it is also common to report mortality rates), but there is extraordinary variation in different types of quality of acute care within the same hospital. In England, for example, surveys have shown different patient experiences for different types of cancer within the same hospital; and CHI’s reviews of clinical governance would often find a dysfunctional clinical team within a well-run acute hospital. So whilst it makes sense to describe performance at the level of the hospital for non-clinical services, such a cleanliness and food, a few indicators of clinical quality will tell us about those selected services only, and not the heterogeneous mix of services that make up acute care.
75. It seems to us that the key step in developing the next generation of hospital reporting is to produce information that disaggregates acute care into different specific services, possibly starting with 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 been by a private insurer (BUPA), which has since 1998, used the Short-Form 36 (SF-36) questionnaire to monitor changes in health status after adult elective surgery. Over 70 independent hospitals across the United Kingdom have collected data on over 100000 patient episodes. Results are reported confidentially, putting the emphasis on supporting a learning culture (Vallance-Owen et al 2004).

7. E-scorecard

76. The E-scorecard is an exciting development and clearly offers great potential for engaging stakeholders in the available information and providing customised outputs. We feel this is a very important area for future work and worthy of considerable time and effort. Here we only comment on some specific additions that might be considered.

77. Elsewhere we have argued for an attractive high-level executive summary for each hospital: this would make a natural addition, with the ability to 'drill-down' being a great advantage of the electronic presentation. Users should then be able to see precisely how their 'scores' for specific areas arose, and to be able graphically to compare their performance with their peers at both an aggregate and specific level
78. We have also 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. In the UK, extensive use is being made of the Geowise software based on AVG graphics to display comparative data: see <http://www.instantatlas.com/health.xhtml> for a range of examples, including the use of funnel plots by the [Association of Public Health Observatories](#). These suggest a variety of developments for multiple views of a common dataset, possibly featuring longitudinal plots, interactive graphics in which specific points are followed through multiple views, and so on. *We recommend the development of attractive interactive interfaces to increase the use and visibility of the information and as a valuable research project.*
79. The E-scorecard currently uses sums of 'reds' and 'greens' as simple summaries of performance: we have argued elsewhere that this suggests paying strong attention to developing a consistent approach to banding.
80. The 'gauge' is an attractive display tool. However the use of confidence intervals is only appropriate when the indicator allows their calculation, and we have discussed elsewhere that a consistent but flexible approach is needed to set tolerances around any benchmark: these tolerances may be constant or reflect size. In each case the interval is placed around the target rather than the observation: this obviates the need to explain confidence intervals, and generalises to situation where the tolerance interval is based on judgement or an empirical distribution of results. See, for example, the website <http://heartsurgery.healthcarecommission.org.uk/>. *We therefore recommend consideration of replacing confidence intervals by tolerance intervals around the target.*

8. Future developments

81. Appendix 5 outlines what those involved in their development and production of *Hospital Reports* see as natural developments and, understandably, as the Mental Health Report is at an early stage, that had a much larger set than the others. As outsiders we support the emphases on the difficult challenges of developing valid benchmarks and a capacity to drill down from headline indicators. We have also argued for working towards a consistent approach to banding (performance classification) and aggregation. We comment here on three other ideas that were not identified by the different *Hospital Reports*.

82. First, the developments currently proposed reflect the way the work on *Hospital Reports* is organised with considerable autonomy and independence across the different teams. Thus, for example, there is interest in understanding relationships between quadrants in each *Hospital Report*, but apparently there is no interest in understanding relationships between the different *Hospital Reports*. *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.*
83. Second, given that the English system of star ratings is so dominated by targets for performance on waiting times, it is frankly a shock to see a system of reporting hospital performance directed at boards of directors and senior managers that has no information whatsoever on waiting times. This raises the question of the purpose of the HRRC: it may be argued that its role is to report on other aspects of performance. But this still seems unsatisfactory to us as directors and senior managers have to look at performance in the round. If, e.g., it were the case that quality was suffering from undue emphasis on reducing waiting times, it would be helpful to have this information brought together in one place by the *Hospital Reports*. *We recommend that HRRC develops Hospital Reports to include information on waiting times.*
84. Third, we understand that in Ontario each hospital may have its own staff survey. This allows for examination of changes over time only and does not offer a basis for benchmarking. This used to be the case in England, where there has been the important development of a standard survey of staff attitudes, which are known to affect organizational outcomes, especially patient care, both directly and indirectly³³. Staff receive a questionnaire that provides information that is organized into ten domains (Healthcare Commission, 2005c)³⁴. *We recommend that HRRC develop a standard staff survey, which can be used in the Hospital Reports and by each hospital to offer a basis for benchmarking.*

³³ For example, staff with positive attitudes towards their work are generally likely to work more effectively and efficiently; staff who feel over worked are more likely to suffer from stress and be away from work as a result, increasing the workload of others; and staff who feel dissatisfied with their job are more likely to leave the organization, which leads to staff shortages and a greater strain on resources. Good management and leadership (as reported by staff) also strongly predict hospital performance.

³⁴ These are: work-life balance; appraisal; training, learning and development; team working; health and safety; errors and incidents; work pressure; staff jobs; management and supervision; extent of positive feeling within organizations; equal opportunities; whistle-blowing, harassment, bullying and violence.

9. Research priorities

85. In this section we make three recommendations, which look to us to be most promising in terms of developing information to have an impact on providers and exploit the opportunities offered by the database that has been generated by producing Reports.
86. HRRC is in essence a technical exercise that aims to transform various kinds of data into useful information. And is very impressive in its scope and methods. But we do not know how this information is actually used within, and across hospitals, so that those with scope to improve can learn from the best. This requires a different kind of expertise from that currently with HRRC. But knowing how information can help local action would also 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. These were identified as problems with the Scottish CRAG Reports and one reason for their lack of impact (Mannion and Goddard, 2001; CRAG, 2002). We have recommended the development of attractive interactive interfaces to increase the use and visibility of the information and as a valuable research project. *More generally 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.*
87. 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: the funnel plot for paediatric cardiac surgery (Figure 4) shows better outcomes from large centres with high volumes; for CCC, however, it may be that small units are better. Whilst it is relatively straightforward to relate size to performance, understanding the drivers of any relationship is more difficult, and is likely to require a combination of quantitative and qualitative research methodologies. *We recommend that HRRC develops research by data mining to examine changes over time and optimal size for different services to generate hypotheses to be explored by qualitative analysis.*
88. Analysis of medical practice from small area variations in rates have been used to identify 'supply sensitive' services³⁵ which largely accounted for the twofold variations in Medicare spending across Regions in the US, and variations in these services resulted in neither more effective care, nor elevated rates of elective surgery, nor better health outcomes (Wennberg et al 2002). Work in England and Wales identified used small area variations to

³⁵ physician visits, specialist consultations, and hospitalisations, particularly for those with chronic illnesses or in their last six months of life.

identify a category of high-variation admissions, for which the only plausible explanation was medical discretion, which accounted for about half of all acute admissions. This category was used to identify discretionary admissions, which are of questionable benefit, across populations in Wales (Fone et al, 2002) and acute hospitals in England (Bevan et al 2004).

We recommend that HRRRC considers using small area variations as a way of making progress in analysis of acute services.

10. Conclusions

89. We were impressed by HRRRC: it is a superb example of a research collaboration, has produced important pioneering *Hospital Reports* in assessing health care performance with wide scope and high technical competence. It is clear that the atmosphere within which HRRRC operates is very different from that of regulators of quality of health care in England: in Ontario the emphasis is on identifying and celebrating success ('naming and faming'); and in England on identifying and penalising failure ('naming and shaming'). This marked contrast obviously means that it would be mistaken to seek simply to impose methods developed for the very different atmosphere in England into Ontario. But this contrast brings out three important issues.
90. The first is the most straightforward: in Ontario and England the overriding interest is on identifying 'outliers': in Ontario the best and in England the worst. For indicators in which chance, uncontrollable variability plays a substantial part, funnel plots offer a sound statistical basis to do this and give a clear graphical portrayal and is easily explained and understood. Other indicators will be more 'controllable'. We have suggested a division of indicators into three 'Types' with corresponding approaches to setting 'tolerances' that define whether an institution is an outlier. For many indicators there is necessarily an element of judgement in setting of such tolerances and it needs to be clearly acknowledged that this is not simply a 'statistical' issue.
91. The second is whether HRRRC ought to move to provide information for the government as part of new developments in accountability, which would mean moving from a 'soft' to a 'hard' approach to performance assessment. We have argued that it is would be difficult for HRRRC to practice both the 'hard' and 'soft' approaches and that HRRRC ought to continue to practice the 'soft' approach, which what it has been, and still is, designed to do, and to continue to act as a stimulus to quality improvement at arms length from government.
92. The third is the question over how HRRRC can best realise its aim 'to facilitate local quality improvement programs and to support hospitals' accountability to the communities they serve' given its primary audiences is 'boards of directors and senior managers'. We are aware of the striking differences between impacts of the annual reports of 'star ratings' in England, and of the Clinical Resource and Audit Group in Scotland: the former had a dramatic impact and the latter little if any. We suggest the HRRRC commissions a study to evaluate the impact of its reports and its value to boards of directors and senior managers.

93. The comparisons between the English and Scottish exercises suggest that for *Hospital Reports* to have an impact, they need to produce information that is: seen to be credible, timely, and covers what is important; easily understood, publicised and widely disseminated; designed to enable benchmarking and learning from the best; and organised to relate to responsibility for improving performance. This raises the question of how HRRC might achieve greater publicity for the *Hospital Reports* to achieve wider dissemination.
94. We have suggested that HRRC develops ways that enable directors and senior managers have a clear overview so that they can which services are the best in Ontario and which services in their own hospital ought to be prioritised for improvement. This raises technical questions of methods of aggregation and of developing ways of drilling down from high level indicators to extract details for those who need to learn and act.
95. The current organisation of *Hospital Reports* seems to reflect their origins and development to extend their scope. The briefing material that was helpfully supplied to us brought out the many differences in methods used between sectors and quadrants. We have identified other important differences in detail in, e.g., key messages and presentation. There is obviously a need for greater consistency across sectors and quadrants. This would help directors and senior managers understand how their hospital performs in different *Hospital Reports*. We suggest, however, also going beyond this to consider changing the organisation of reports to bring together different services on a hospital basis and including information on hospital waiting times.
96. A much bigger question here, which we believe is central to the next generation of *Hospital Reports*, are developments to produce information on specific services within acute care. We have suggested two possible new kinds of data that could be collected routinely: a survey of staff and a questionnaire on health status following discharge from hospital.
97. We have suggested a number of areas for research and development: methods of determining benchmarks as alternative to statistical distribution; setting tolerance limits to reflect practical rather than statistical significance; developing ways of reporting information to ease this being used to improve performance; examining changes over time; and investigating the optimal size for different kinds of services.

11. References

Audit Commission (2003) *Waiting list accuracy*. London: The Stationery Office. <http://www.audit-commission.gov.uk/reports/NATIONAL-REPORT.asp?CategoryID=english^574&ProdID=2BBB4A89-2212-401A-B385-02CA01D10100>

Bevan, G and Hood, C (2006a) Have targets improved performance in the English NHS? *BMJ*, 332, 419-422.

Bevan, G and Hood, C. (2006b) What's Measured is What Matters: Targets and Gaming in the English Public Health Care System. *Public Administration* **84**(3): 517-38.

Bevan G. (2006) Setting Targets for Health Care Performance: lessons from a case study of the English NHS. *National Institute Economic Review* 197: 67-79.

Bevan G, Hollinghurst S, Benton P, Spark V, Sanderson H, Franklin D (2004) Using information on variation in rates of supply to question professional discretion in public services. *Financial Accountability and Management*, **20** (1): 1-17.

Carter, Neil, Klein, Rudolf and Day, Patricia (1995) *How Organisations Measure Success. The use of performance indicators in government*. London: Routledge.

Commission for Health Improvement (2003a) *NHS performance ratings. Acute trusts, specialist trusts, ambulance trusts 2002/03*. London: The Stationery Office.
<http://ratings2003.healthcarecommission.org.uk/ratings/>

Commission for Health Improvement (2003b) *NHS performance ratings. Primary care trusts, mental health trusts, learning disability trusts 2002/03*. London: The Stationery Office.
<http://ratings2003.healthcarecommission.org.uk/ratings/>

Commission for Health Improvement (2003c) *What CHI Has Found in: Ambulance Trusts*. London: The Stationery Office.

<http://www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/Ambulance/fs/en>

Commission for Health Improvement. 2004. *What CHI Has Found in: Acute Services*. London: The Stationery Office.
<http://www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/AcuteAndSpecialist/fs/en>.

CRAG (2002) *Clinical Outcome Indicators. A report from the Clinical Outcomes Working Group.*
www.show.scot.nhs.uk/crag

Department of Health (1998) *A First Class Service: Quality in the New NHS.* London: Department of Health.

http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4006902&chk=j2Tt7C.

Department of Health (2001) *NHS performance ratings acute trusts 2000/01.* London: Department of Health.

http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4003181&chk=wU4Zop

Department of Health (2002a) *NHS performance ratings acute trusts, specialist trusts, ambulance trusts, mental health trusts 2001/02.* London: Department of Health.

http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4002706&chk=dBD1wB

Fone D, Hollinghurst S, Bevan G, Coyle E, Palmer S. (2002) Information for clinical governance: analysis of routine hospital activity data for Wales. *Journal of Public Health Medicine*, **24** (4): 292-98.

Healthcare Commission (2004) *2004 performance ratings.* London: Healthcare Commission. <http://ratings2004.healthcarecommission.org.uk/>

Healthcare Commission (2005a) *NHS performance ratings 2004/2005.* London: Healthcare Commission. <http://ratings2005.healthcarecommission.org.uk/>

Health Care Commission (2005b) *Assessment for improvement. The annual health check.* London: Health Care Commission.

http://www.healthcarecommission.org.uk/InformationForServiceProviders/AnnualHealthCheck/fs/en?CONTENT_ID=4017483&chk=ub2qrx.

Healthcare Commission (2005c) *NHS national staff survey 2004*

Summary of key findings. London: Healthcare Commission.

<http://www.healthcarecommission.org.uk/assetRoot/04/01/66/69/04016669.pdf>

Hibbard JH, Stockard J, Tusler M. (2003) Does publicizing hospital performance stimulate quality improvement efforts? *Health Affairs* **22**(2): 84-94.

Kornai, Janos (1994) *Overcentralisation in economic administration.* Oxford: Oxford University Press.

Mannion, Rand Goddard, M (2001) Impact of published clinical outcomes data: Case study in NHS hospital trusts. *BMJ*, **323**: 260-263.

Marshall, Martin N., Shekelle, Paul G., Davies, Huw T. O., and Smith, Peter C. (2003) Public reporting on quality in the United States and the United Kingdom. *Health Affairs*, **22**(3): 134–148.

National Audit Office (2001) *Inappropriate adjustments to NHS waiting lists* (HC 452) London: The Stationery Office. www.nao.gov.uk/publications/nao_reports/01-02/0102452.pdf

Nove, Alec (1958) The Problem of Success Indicators in Soviet Industry, *Economica* (New Series) **25** (97): 1-13.

Scally, G., and L.J. Donaldson (1998) Looking forward: Clinical governance and the drive for quality improvement in the new NHS in England, *BMJ* **317**: 61 - 65.

Smith, P (1995) On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, **18**: 277–310.

Wennberg JE, Elliott SF, Skinner JS. (2002) Geography and the debate over Medicare reform. *Health Affairs*, Web Exclusive, February 13, (19 pages).

Woodham-Smith C (1970) *Florence Nightingale*. London: Fontana, 230.

Appendix 1: Schedule of Visit by Bevan and Spiegelhalter

Wednesday, April 12th

9:00 – 9:30 Louise Lemieux-Charles and Geoff Anderson (456)

9:30 – 10:30 Neil, Carey and Geoff [Jenn Wagg] (Room 451)

- SIC data and measurement issues

10:30 – 11:30 Steini, Paula, Carey and Geoff (Room 451)

- Overview of Hospital Report, methodological Issues and goals for visit

12:00 – 2:00 Gwyn, David, Geoff, Michael, Carey [Heather Dawson, Jeanie Lacroix, Eugene Wen] (Room 451 – lunch ordered in)

- Lunch and discussion of profiling/ categorization of performance by indicator

2:00 – 3:00 Michael and Carey [Marcus Loreti] (Room 451)

- Patient satisfaction data and measurement

3:00 – 4:00 Janet, Betty, Scott [Jenn Wagg] (Room 451)

- Mental Health

4:00 – 5:00 Geoff, Asma and Carey [Jenn Wagg] (Room 451)

- Clinical data from ED and acute care plus first day wrap up

Thursday, April 13th

9:00 – 10:00 Carey, Ian, George (will be phoning in), Imtiaz and Linda (HRRC Mtg Rm)

- Finance data and measurement

10:00 – 11:00 Walter, Gary (will be phoning in), Geoff and Carey (HRRC Mtg Rm)

- Complex Continuing Care

11:00 – 12:00 Cheryl, Susan, Geoff and Carey (Room 451)

- Rehab

12:00 – 2:00 Gwyn, David, Geoff, Carey [Heather Dawson, Jeanie Lacroix, Eugene Wen]
(Room 451 – lunch ordered in)

- Lunch and discussion of benchmarking and indices of performance

2:00 – 3:00 Carey and Jillian [Nicole Howe] (Room 451)

- Women's Health

3:00 – 4:00 Steini, Paula, Geoff and Carey (Room 451)

- Wrap up and next steps

Appendix 2: List of CRAG's Clinical Outcome Indicators

Pregnancy under the age of 16

Therapeutic abortion rates (s)

Childhood incidence of measles

Cervical cancer mortality

Suicide rate

Rate of emergency admission for diabetic ketoacidosis

Longer in-patient stays for children with asthma

30 day survival after admission for fractured neck of femur

Discharge home within 56 days of admission with hip fracture

30 day survival after admission for acute myocardial infarction

Re-operation within 1 year of transurethral prostatectomy

Emergency re-admission within 28 days of discharge from medical specialty

30 day survival after admission for stroke

Discharge home within 56 days of admission for stroke

Psychiatric inpatients: death within 1 year of discharge

Psychiatric inpatients aged 65+: death within 1 year of discharge

Psychiatric inpatients: suicide within 1 year of discharge

Proportion of first births by caesarean section

Vaginal delivery after caesarean section

Babies admitted to a neonatal unit (s)

28 day emergency re-admission: removal of tonsils/adenoids

D & C rates in women under 40

Use of medical methods for early termination of pregnancy

Survival with cancer of the trachea, bronchus and lung

Survival with cancer of the large bowel

Breast cancer (s)

Survival with cancer of the ovary

28 day emergency re-admission: elective operation for cataract

28 day emergency re-admission: emergency appendectomy

28 day emergency re-admission: elective prostatectomy

28 day emergency re-admission: elective hysterectomy

28 day emergency re-admission: elective total hip replacement

Survival with cancer of the stomach

Survival with cancer of the cervix uteri

Cardiac procedures - standardised procedure ratios for coronary

angiography, angioplasty and CABG (s)

Breast feeding

Smoking during pregnancy

Registration with general dental practitioner (s)

Decayed, Missing and filled teeth in children age 5 years (s)

Colorectal cancer (s)

Emergency admissions (s)

Primary Care Indicators: Prescribing and Immunisation rates (s)

Mortality within 30 days of elective surgery

Emergency readmission rates within 7 and 28 days of discharge

Alcohol problems (s)

Appendix 3: Lists of Indicators in Key Targets and the Balanced Scorecard by Type of Organization for Star Ratings Published in 2003

Acute

Key Targets (9)

Number of in-patients waiting longer than the standard
Number of out-patients waiting longer than the standard
A&E emergency admission waits (12 hours)
Total time in A&E
Cancelled operations not admitted within 28 days
Two-week cancer waits
Improving working lives
Hospital cleanliness
Financial management

Clinical Focus (10)

Clinical negligence
Deaths within 30 days of a heart bypass operation
Deaths within 30 days of selected surgical procedures
Emergency readmission to hospital following discharge
Emergency readmission to hospital following discharge for children
Emergency readmission to hospital following treatment for a stroke
Emergency readmission to hospital following treatment for a fractured hip
Infection control procedures
Methicillin Resistant Staphylococcus Aureus (MRSA) bacteraemia: improvement score
Thrombolysis treatment time

Appendix 3: (continued)

Acute (continued)

Patient Focus (19)

Six-month in-patient waits

Total in-patient waits

Thirteen-week outpatient waits

Day-case booking

A&E emergency admission waits (4 hours)

Cancelled operations

Nine-month heart operation waits

Waiting times for Rapid Access Chest Pain Clinic

Breast cancer treatment

Delayed transfers of care

Out-patient/A&E survey - Better information, more choice

Out-patient/A&E survey - Clean, comfortable, friendly place to be

Out-patient/A&E survey - Building relationships

Out-patient/A&E survey - Safe, high-quality, co-ordinated care

Out-patient/A&E survey - Access and waiting

Pediatric outpatient did not attend rates

Patient complaints procedure

Better hospital food

Privacy and dignity

Capacity and Capability (7)

Data quality

Staff opinion survey

Junior doctors' hours

Consultant appraisal

Sickness absence rate

Information governance

Fire, health, and safety

Appendix 3: (continued)

Mental Health

Key Targets (7)

Assertive Outreach Team Implementation
Care Programme Approach integration
Mental Health Minimum Dataset implementation
Number of out-patients waiting longer than the standard
Improving working lives
Hospital cleanliness
Financial management

Clinical Focus (5)

Clinical negligence
Care Programme Approach systems implementation
Psychiatric readmissions (Adult)
Psychiatric readmissions (Older people)
Suicide rate

Patient Focus (6)

Transition of care between mental health services for adults and for older people.
Transition of care between mental health services for children and adolescents for and adults
Patients with copies of their own care plan
Patient complaints procedure
Better hospital food
Privacy and dignity

Appendix 3: (continued)

Mental Health

Capacity and Capability (12)

Missed out-patient appointments

Crisis Resolution Team Implementation

Out of catchment area treatments (Adults)

Out of catchment area treatments (Older people)

Mapping of mental health services for children and adolescents

Data quality

Staff opinion survey

Junior doctors' hours

Consultant appraisal

Sickness absence rate

Information governance

Fire, health, and safety

Appendix 3: (continued)

Ambulance Trusts

Key Targets (4)

Category-A calls meeting 8-minute target

Category-A calls meeting 14/19-minute target

Improving working lives

Financial management

Clinical Focus (2)

Clinical negligence

Thrombolysis protocols and procedures: Training of paramedic staff

Patient Focus (3)

Category-B/C calls meeting national 14/19-minute target

GP urgent calls meeting national 15-minute target

Patient complaints procedure

Capacity and Capability (4)

Staff opinion survey

Sickness absence rate

Information governance

Fire, health, and safety

Appendix 3: (continued)

Primary Care Trusts (PCTs)

Key Targets (9)

Access to a general practitioner (GP)
Access to a primary care professional (PCP)
Number of in-patients waiting longer than the standard
Number of out-patients waiting longer than the standard
Total time in A&E
Single telephone access – implementation plans
Four-week smoking quitters
Improving working lives
Financial management

Access to Quality Services (14)

Emergency readmission to hospital following treatment for a fractured hip
Percentage of general practitioner (GP) practices in a shared care scheme
Sexual health: Access to services for early unintended pregnancy
Level of 24-hour access to specialist mental health services
A&E emergency admission waits (12 hours)
Twelve-month heart operation waits
Delayed transfers of care
Access to NHS dentistry
Primary Care Trust (PCT) survey - Access and waiting
PCT survey - Better information, more choice
PCT survey - Building closer relationships
PCT Survey - Clean, comfortable, friendly place to be
PCT Survey - Safe, high quality, coordinated care
Prescribing of atypical antipsychotics

Appendix 3: (continued)

Primary Care Trusts (PCTs)

Improving Health (10)

Death rates from circulatory diseases (change in rate)

Deaths rates from accidents, all ages (change in rate)

Death rates from cancer age <75 (change in rate)

Breast cancer screening

Cervical cancer screening

Flu vaccinations

Teenage pregnancy: Conceptions below age 18 (change in rate)

Diabetes services baseline assessment

Coronary Heart Disease Audit

Suicide audit

Service Provision (13)

Emergency admissions (change in rate)

Emergency admission to hospital for children with lower respiratory tract (LRT) infections
(change in rate)

Primary care management - acute conditions (change in rate)

Primary care management - chronic conditions (change in rate)

Community equipment

Patient complaints procedure

Prescribing of antibacterial drugs

Prescribing rates for drugs acting on benzodiazepine receptors

Staff opinion survey

General practitioner (GP) appraisal

Sickness absence rate

Fire, health, and safety

Generic prescribing

Appendix 4: Comments on specific sectors and quadrants

This Appendix gives detailed examples of the need for more consistency in performance classification, and more careful consideration of whether facility size should influence the critical thresholds used in performance classification. Although most of the examples are for Acute Care, and many of these comments apply across other sectors.

Acute Care

- i. p 9 etc. "Statistical significance" based on 99% intervals for patient satisfaction and CUO, in contrast to 95% in other sectors.
- ii. p11 etc. **SIC** based on complex aggregation process from survey responses. Although there has been a substantial effort in constructing the indices, it is difficult from the final presentation to identify what is really driving the conclusion. There is currently a transformation to 'normality' and 1.645 SDs used as a cut-off, which we would expect to be approximately a 90% interval and hence lead to about 5% being each of high/low, which seems to be about what happens. No account is taken of facility size, which appears reasonable. It may be better for the critical thresholds to be explicitly judgemental, in the manner of finance, or be based on previous years' distributions.
- iii. p19 etc **Patient Satisfaction** shows many high/low hospitals. This is presumably because the 99% intervals are based on a fairly precise within-hospital SE and there is substantial over-dispersion of the indicator around the overall mean. Some allowance for over-dispersion appears appropriate, as is used by the Healthcare Commission when banding patient surveys: essentially the bands are based on both the within-hospital and between-hospital variability.
- iv. p31 etc. The **CUO** data only reports percentages and information on numerators and denominators would be helpful, if only displayed in a funnel plot. This would presumably help explain why some apparently extreme results (eg 12.1% for re-admissions in the Hospital for Sick Children) was not 'significant'.
- v. p37 etc The Table of results for show no high/low values for **Finance**, as judgemental thresholds for Total Margin and Current Ratio are only in the e-scorecard. Extension to all the indicators seems appropriate.
- vi. p46 etc In the **Women's health** section, some funnel plots are used but without the funnels! The use of $(F-M)/F$ can lead to some rather odd results if F is low eg readmission for congestive heart failure in Glengarry Hospital has $F=7.9$, $M=35.8$, so that $(F-M)/F = -3.54$, which is difficult to interpret. Technically one might want to use $\log(\text{odds ratios})$, (which are even more uninterpretable), but I wonder if simple M/F would be clearer.
- vii. For Women's health, if expected counts of adverse events are less than 2 then currently a hospital is not included. This is reasonable if only concerned with good performance (as can never be shown to be doing very well), but is not appropriate if also want to indicate poor performance.

Rehabilitation

- i. p13 etc. For **SIC**, a high/low classifications based on 99% interval for the provincial mean gives many signals and does not really seem an appropriate basis. See comments on SIC under Acute Care.
- ii. p17 etc. For **CUO**, high/low based on 95% intervals gives many signals, suggesting over-dispersion. As there is interest in size of facility, funnels seem particularly appropriate.

Complex Continuing Care

- i. For **SIC**, see comments from Acute Care
- ii. Both patient satisfaction and CUO make many exclusions based on sample size, and also use the 'low cut-off' rule to prevent large sizes being penalised. A consistent approach to relating thresholds to size would help.

Emergency Department

- i. p9 It could be considered whether the very different numbers of hospitals identified as 'high-performing' in the different quadrants is reasonable or due to the methods used within each quadrant.

Mental Health

- i. Comparisons are made on a regional basis rather than individual hospitals. There are cogent reasons for this, but a move towards display at a finer level would be welcome.
- ii. It is surprising to see LOS in SIC, which does not occur in other sectors.

Appendix 5: Areas for R &D identified by Report s

The different *Hospital Reports* identified research and development as follows.

*Acute Care*³⁶

- i. development of CUO indicators for children, chest pain in adults, and management of stroke and transient ischemic attacks (TIA); and
- ii. development of weighting of visits based on resource demands to support inclusion of indicators of efficiency.

*Emergency Department Care*³⁷

- i. analysis of impact of hospital size and community served on Patient Satisfaction³⁸;
- ii. analysis of inter-quadrant relationships³⁹;
- iii. development of new measures to determine improvements in transition from acute care to other levels of care, including home⁴⁰;
- iv. analysis of issues of equity between the care of men and women.

*Rehabilitation*⁴¹

- i. development of the scope covered to include acute care, outpatient ambulatory and home; and
- ii. research into data quality, grouping and associated weights for Financial Performance and Condition.

*Complex Continuing Care*⁴²

³⁶ Areas for future study (page 4)

³⁷ Developments for the next Report to be released in 2007 (page 4)

³⁸ The Report observes that 'Many Toronto (and Greater Toronto Area) hospitals scored below-average in the Patient Satisfaction quadrant, while 50% of small hospitals met the criteria for high performing hospitals. Across all indicators, and in both fiscal years, on average, small hospitals scored higher than community and teaching hospitals'.

³⁹ to help to determine which indicators most significantly impact on quality of care, patient experience and outcomes

⁴⁰ from both the hospital and patient perspective

⁴¹ Developments for future Reports (page 4)

⁴² Next steps (page 4)

- i. development of new indicators relating to the use and management of human resources⁴³;
- ii. refinement of risk-adjustment of patient satisfaction and clinical indicators;
- iii. research into valid benchmarks;
- iv. development of methods for hospitals to “drill down” from Hospital Report indicators;
- v. research on valid methods to capture patient satisfaction information from short-stay CCC patients⁴⁴;
- vi. evaluation of the uptake and effectiveness of quality improvement initiatives⁴⁵.

*Mental Health*⁴⁶

- i. Development of data sources⁴⁷:
 - more specific information by refining SIC survey questions to develop more precise indicators that can be “drilled down” to a level of detail that is actionable;
 - improving data sources for the SIC survey so that reporting is based on aggregated client data rather than best estimates of program practice;
 - seeking better measures of patient severity;
 - revising the definition of Alternative Level of Care (ALC) to be more broadly relevant across the system;
 - including the remaining non-divested provincial psychiatric hospitals in the calculation of client-level indicators;
 - continuing to explore the feasibility of obtaining valid, system-wide patient outcome information;
 - continuing developmental work on the patient perception of care survey.
- ii. Refining Indicators
 - Continuing work on indicators under development⁴⁸;
 - Reviewing the relevance of the indicators to hospitals’ strategic priorities⁴⁹.

⁴³ for Financial Performance and Condition and SIC quadrants

⁴⁴ who are currently under-represented relative to their total numbers using the current patient satisfaction survey approach

⁴⁵ to understand whether process changes were made in the short-term, and are leading to improved outcomes in the long-term

⁴⁶ Next steps (pages 94-95)

⁴⁷ And also supporting the many voices across the province requesting improvement of the infrastructure necessary to create and maintain decision-support information systems as the information sought was often routinely gathered but was not stored or managed in a way that would facilitate retrieval.

⁴⁸ including client outcome, a global measure of appropriateness of care, and perception of care indicators

iii. Interpretation of Indicators

- continuing to test approaches to facilitate interpretation of indicators⁵⁰;
- developing program groupings within specialty facilities and considering how this will be reflected in performance indicators;
- working with HRRC to develop benchmarking approaches or other means of assessing performance.

iv. Scope of Reporting

- extending the scope of the Report to include non-Schedule 1 hospitals, and hospital outpatient programs; and
- adding community mental health and other services⁵¹.

⁴⁹ using feedback from hospitals about their individual mental health reports, and complementary work by HRRC to identify hospitals' strategic priorities

⁵⁰ such as: cross-quadrant analyses; triangulation of data sources to assess an issue; expansion of contextual variables; stratification and risk adjustment of hospital-specific results

⁵¹ in the longer term, as data sources become available