J. R. Statist. Soc. A (2005)
168, *Part* 1, *pp.* 1–27

Performance indicators: good, bad, and ugly

[The report of a Working Party on Performance Monitoring in the Public Services chaired by Professor S. M. Bird, submitted on October 23rd, 2003]

Membership of the Working Party

Sheila M. Bird (*Chair*) (Medical Research Council Biostatistics Unit, Cambridge, and Department of Statistics and Modelling Science, University of Strathclyde)
Sir David Cox FRS (Nuffield College, Oxford)
Vern T. Farewell (Medical Research Council Biostatistics Unit, Cambridge)
Harvey Goldstein FBA (Institute of Education, University of London)
Tim Holt CB (Department of Social Statistics, University of Southampton)
Peter C. Smith (Department of Economics and Related Studies, University of York)

Summary. A striking feature of UK public services in the 1990s was the rise of performance monitoring (PM), which records, analyses and publishes data in order to give the public a better idea of how Government policies change the public services and to improve their effectiveness.

PM done well is broadly productive for those concerned. Done badly, it can be very costly and not merely ineffective but harmful and indeed destructive.

Performance indicators (PIs) for the public services have typically been designed to assess the impact of Government policies on those services, or to identify well performing or underperforming institutions and public servants. PIs' third role, which is the public accountability of Ministers for their stewardship of the public services, deserves equal recognition. Hence, Government is both monitoring the public services and being monitored by PIs.

Especially because of the Government's dual role, PM must be done with integrity and shielded from undue political influence, in the way that National Statistics are shielded. It is in everyone's interest that Ministers, Parliament, the professions, practitioners and the wider public can have confidence in the PM process, and find the conclusions from it convincing.

Before introducing PM in any public service, a PM protocol should be written. This is an orderly record not only of decisions made but also of the reasoning or calculations that led to those decisions. A PM protocol should cover objectives, design considerations and the definition of PIs, sampling *versus* complete enumeration, the information to be collected about context, the likely perverse behaviours or side-effects that might be induced as a reaction to the monitoring process, and also the practicalities of implementation. Procedures for data collection, analysis, presentation of uncertainty and adjustment for context, together with dissemination rules, should be explicitly defined and reflect good statistical practice. Because of their usually tentative nature, PIs should be seen as 'screening devices' and not overinterpreted. If quantitative performance targets are to be set, they need to have a sound basis, take account of prior (and emerging) knowledge about key sources of variation, and be integral to the PM design.

Aspirational targets have a distinctive role, but one which is largely irrelevant in the design of a PM procedure; motivational targets which are not rationally based may demoralize and distort. Anticipated and actual side-effects of PM, including on individuals' behaviour and priorities, may need to be monitored as an intrinsic part of the PM process.

Independent scrutiny of PM schemes for the public services should be set up and must report publicly. The extent and nature of this scrutiny should be related to the assessed drawbacks and benefits, reflect ethical concerns, and conform with good statistical practice.

Research is needed into the merits of different strategies for identifying institutions or individuals in the public release of PM data, into how new PM schemes should be evaluated, and into efficient designs for evaluating a series of new policies which are monitored by PIs.

The Royal Statistical Society considers that attempts to educate the wider public, as well as policy makers, about the issues surrounding the use of PIs are very important. High priority should be given to sponsoring well-informed public debate, and to disseminating good practices by implementing them across Government.

 0964–1998/05/168001

Specific recommendations

1. All PM procedures need a detailed protocol (Section 2).
2. A PM procedure must have clearly specified objectives, and achieve them with methodological rigour (Sections 2 and 3). Individuals and/or institutions monitored should have substantial input to the development of a PM procedure.
3. A PM procedure should be so designed that counter-productive behaviour is discouraged (Sections 2, 5 and 7).
4. Cost-effectiveness should be given wider consideration in both the design and the evaluation of PM procedures. Realistic assessment of the burden (indirect as well as direct) of collecting quality-assured PM data is important, for PM's benefits should outweigh the burden (Sections 2, 3, 5 and 6).
5. Independent scrutiny of a PM procedure is needed as a safeguard of public accountability, methodological rigour, and of the individuals and/or institutions being monitored. The scrutineers' role includes checking that the objectives of PM are being achieved without disproportionate burden, inducement of counter-productive behaviours, inappropriate setting or revision of targets, or interference in, or overinterpretation of, analyses and reporting (Sections 2, 3, 4 and 7).
6. PIs need clear definition. Even so, they are typically subject to several sources of variation, essential or systematic—due to case mix, for example—as well as random. This must be recognized in design, target setting (if any) and analysis (Sections 2 and 3).
7. The reporting of PM data should always include measures of uncertainty (Section 4).
8. Investigations on a range of aspects of PM should be done under research council sponsorship, including study of the relative merits of different dissemination strategies for the public release of PM data (Section 6).
9. Research should also be undertaken on robust methods for evaluating new Government policies, including the role of randomized trials. In particular, efficient designs are needed for when Government departments, in accordance with budgetary or other constraints, introduce (or 'roll-out') a series of PI-monitored policies (Section 6).
10. Ethical considerations may be involved in all aspects of PM procedures, and must be properly addressed (Sections 2, 3, 4, 5 and 7).
11. A wide-ranging educational effort is required about the role and interpretation of PM data (Section 7).

1. Introduction

1.1. Background to performance monitoring in the public services

A striking feature of UK public services in the 1990s was the rise of performance monitoring (PM). PM was introduced across Government in an attempt to measure the processes and outcomes of the public services, and as a goad to efficiency and effectiveness. At the same time, PM provided clearer accountability to Parliament and the public for the Government's stewardship of the public services. Such monitoring can take place at the level of an entire programme (e.g. the impact of the National Health Service (NHS) on the nation's health), an organization delivering a service (e.g. a police authority), or an individual (e.g. a surgeon). The rise of PM was linked to the following: to the increased capacity to record aspects of the public services which has been brought about by developments in information technology; to demands for increased accountability both of public services and professionals (Bristol Royal Infirmary Inquiry Panel, 1999, 2001; Spiegelhalter *et al.*, 2002; Smith, 2001); to greater use of explicit contracts for the provision of public services; and to heightened pressures to ensure that public finances are spent efficiently.

Three broad aims of public service performance data can be identified: to establish 'what works' in promoting stated objectives of the public services; to identify the functional competence of individual practitioners or organizations; and public accountability by Ministers for their stewardship of the public services. These can be viewed, respectively, as the research, managerial and democratic roles of performance data. In each role, PM data can be used to

promote change, by enabling policy makers to adopt the most cost-effective technologies, by helping practitioners to identify opportunities for personal or organizational improvement, and by allowing voters to judge the real impact of the Government's policies. Ministerial resignation notwithstanding, the second use of performance data has predominated—examining the performance of, and scope for improvement in, the providers of public services. This is not surprising since the use of administratively generated data for research purposes, as in the first objective above, is widely regarded as problematic. Also for PM of individuals or institutions, their suitability needs critical consideration (Spiegelhalter *et al.*, 2002; Dyer, 2003; Carter, 2003; Jacobsen *et al.*, 2003; Aylin *et al.*, 2003; Bridgewater *et al.*, 2003).

PM in the public services has a number of stakeholders. Professionals and others responsible for delivering public services might use PM data to check on their own performance and to seek out scope for improvement. Managers might use PM data to check that their organizations are delivering services in line with intentions, to set targets, and to act as a basis for performance rewards. Non-executive board members of public services might use PM data to hold managers to account. It is also intended that members of the public could use PM data in a number of ways, e.g. to check that their local services are performing satisfactorily, to inform their voting choices, or to choose a public service provider. And national and local governments might wish to use the data to inform a variety of decisions and political actions.

The systematic use of PM in the UK public services began in the 1980s, but only assumed central importance in the early 1990s, when the Citizen's Charter and its offshoots, such as the Patient's Charter and the Parent's Charter, were developed as instruments for seeking to secure some very minimal standards in selected public services (`http://www.servicefirst.gov.uk/index/list.htm`). Other PM schemes developed in the 1990s include those specifically aimed at local authorities. The PM philosophy reached its apotheosis with the development of national public service agreements and, after the 2001 general election, the UK Government set up a central Office of Public Services Reform with the task of overseeing all aspects of monitoring of public services (`http://www.pm.gov.uk/output/page270.asp#3`). In 2003, the Prime Minister's Delivery Unit was relocated to the Treasury (House of Commons Public Administration Select Committee, 2003a).

Thus, a central feature of Government policy for public service change is its reliance on the use of performance indicators (PIs) to determine whether policy objectives are being met. Indicators are based principally upon routinely collected data that measure processes and outputs or on special surveys such as for consumer satisfaction or drug-related acquisitive crime. In our report we focus on PIs and the issues raised by them, but recognize that other forms of information, notably from on-site inspections (Fairweather (2002); see also `www.chi.nhs.uk`), and also financial data are important for understanding the delivery of a public service.

The UK is not alone in the use of PM. In the United States, many state education authorities, such as California and Texas, use school tests as performance monitors and examples from health and other services are also found. In some countries, however, there has been more general scepticism. Thus New South Wales in Australia has legislated against the use of school league tables based upon test scores, as has the Republic of Ireland. Within the UK, the publication of school league tables has been abolished in Northern Ireland and in Wales. It appears that dissatisfaction with the lack of contextualization (see Section 3) and the negative 'side-effects' (see Section 5) have been important factors in influencing public and political opinions.

1.2. Public Administration Select Committee

In the spring of 2003, the UK's PM of the public services came under scrutiny through hearings of the Public Administration Select Committee (House of Commons Public Administration

Select Committee, 2003b) to which in particular the Statistics Commission submitted written evidence (Statistics Commission, 2003a). The Select Committee made site visits and also interviewed Ministers, senior civil servants and representatives or practitioners from local authorities, prisons, police, teaching and health care as well as from business, trade unions and the Consumers' Association (Consumers' Association, 2003). Its minutes and memoranda (Audit Commission, 2003a; House of Commons Public Administration Select Committee, 2002a, b, 2003c, d) contain a wealth of detail about past attempts at PM, including accounts of perverse behaviours which had been anticipated by practitioners but were ignored in the design of PIs, and of jarring targets. The choice and cascade of indicators and targets exercised the Public Administration Select Committee. These are issues to which we return (see Section 2). The Scottish Executive's initiative to publish together all its public service agreement indicators, technical notes, targets and yearly out-turns was commended; England was to have a similar site established in 2003.

1.3. Royal Statistical Society Working Party's remit

In November 1995, Goldstein and Spiegelhalter read a seminal paper to the Royal Statistical Society on league tables and their limitations (Goldstein and Spiegelhalter, 1996). It was widely discussed, and is a key reference. In subsequent years, the use of PM increased, but still without sufficient heed to the risks that had been outlined of overinterpretation in the presence of large, often inadequately reported uncertainty.

In January 2003, the Royal Statistical Society held a discussion meeting on PIs (Best and Day, 2004), which highlighted the extent to which design, analysis and dissemination of PIs in the public services can bypass the safeguards of National Statistics (Her Majesty's Government, 1998, 1999; Office for National Statistics, 2000, 2002), which a previous Royal Statistical Society Working Party had sought to strengthen (Moore, 1991). Accordingly, in April 2003, the Royal Statistical Society's Working Party on Performance Monitoring in the Public Services was constituted by Council.

By long tradition, from 'passionate statistician' Florence Nightingale in the 19th century (Spiegelhalter, 1999) to recent Presidential addresses such as 'Mad cows and ecstasy: chance and choice in an evidence-based society' (Smith, 1996), the Royal Statistical Society supports the application of scientific method and insightful analysis to improve the condition of society, and to evaluate objectively Government policies. The remit of the Royal Statistical Society's Working Party was therefore to deal with scientific, and in particular statistical, aspects of many issues in PM which the Public Administration Select Committee has highlighted (House of Commons Public Administration Select Committee, 2003b), and to make recommendations.

We seek to present a practical perspective on PM that can help to resolve critical issues in the design, analysis and reporting of PIs in the public services.

Although much could be said about the PM of university research and teaching (Roberts, 2003; Higher Education Funding Council for England, 2002) we have refrained from discussing these issues at length on the general principle that criticisms from those directly involved carry less force than external comment. Nevertheless, across the public services, it is essential that criticisms and concerns from those being, or to be, monitored are seriously addressed.

2. Performance monitoring design, target setting and protocol

PM's three broad but diverse aims (finding out 'what works', identifying the functional competence of practitioners or institutions, and for public accountability) profoundly influence its

design through to reporting. In particular, the choice of unit of study (practitioner or institution, for example) should be consistent with PM's specific aim. For the first and third aims, it is not necessary to collect data from all institutions in order to make inferences; within representative institutions, it is a matter of detailed design whether information is collected about all, or an appropriate sample of, practitioners and their clients (e.g. patients, pupils or prisoners). For the second aim of identifying the functional competence of institutions, all must be monitored and the burden of data collection per institution (Berkshire, 2001) increases significantly. Burden also increases with the number of assessed PIs.

In essence, the following questions have to be answered in setting up a scheme of PM: what is the purpose of PM; what is the unit of study; what should be measured; for reasons of cost, convenience or burden, have proxy indicators been substituted; how should the data be collected; how will the data be checked and audited; how should the data be analysed, including how frequently and adjustment for context; how should the results be presented, including the dissemination rules according to which institutions or individuals may be named publicly; and how will uncertainty be conveyed to safeguard against over-interpretation and misinterpretation?

Answers to these questions about design, analysis and presentation, on which we comment in the following sections, should be documented in a PM protocol, which also explains if, and how, the PM scheme will be augmented by follow-up inspection (Audit Commission (2003b); see also `www.chi.nhs.uk`). The issue of action (Deming, 1986; Department of Health, 2002, 2003; Hawkes, 2003) on the basis of PM information and from follow-up inspections we regard as outside our discussion except in so far as the setting of specific numerical targets is involved (see Sections 2.3 and 2.4).

2.1. Choosing what to measure: sampling versus complete enumeration, potential users and primary objectives

A general approach to the choice of indicators is to identify some broad dimensions of performance that are important. These should not be too numerous (Atkinson *et al.*, 2002) because, typically, several PIs may be needed per dimension for each dimension to be properly reflected.

2.1.1. Example

In transplantation, the dimensions might include measures of transplant survival, numbers of patients transplanted, risk profile of transplantees, patient satisfaction, waiting time for transplantation, and deaths on the transplant waiting list.

Then, for each dimension, there may be several more detailed measures (according to the organ transplanted, say) which, for some but not all purposes, will be aggregated at analysis. It then needs to be specified how the individual detailed measures, patient satisfaction say, are to be obtained.

One key choice is whether sampling or complete enumeration is to be used. In the former case, a clearly specified sampling procedure is essential and must be subject to audit. In either case, checks on data quality are required (Deming, 1986). Note that, in suitable cases, not only may sampling be much less intrusive than complete coverage, but also the quality of the data obtained may be higher, and the expense and predictability of data collection may be considerably reduced.

The potential users of the measures need to be clear. They may be one or more of the following: individuals directly involved in the activity under study, local managers, regional or

central managers, the Government and the general public. In principle, PM gives the public a transparent way to assess the work of Ministers and appointed officers. PM also presents individuals and managers with new, ideally much-wanted, data together with regionally (Bridgewater *et al.*, 2003) or nationally comparative information that was previously unavailable to them, unaffordable or insufficiently analysed (Carter, 2003).

Next, the primary objectives need clarification. They may range from the isolation of potentially badly performing units within a system that is broadly satisfactory to the encouragement of a general improvement of performance, possibly over a short timescale, but often more realistically and fruitfully for improvement to be sustained over a longish period.

2.1.2. Example

A short timescale may be reasonable if robust evidence about the efficacy and cost-effectiveness of a new treatment has led the National Institute for Clinical Excellence to recommend its use by the NHS. Patient outcomes are likely to improve over a short timescale if uptake of the new treatment is rapid throughout the NHS. However, over a 4-year period, it was difficult to identify consistently improving (or deteriorating) schools (Gray *et al.*, 2001).

2.1.3. Example

On the other hand, if there is a dearth of research evidence about what works, for example in reducing reoffending, then improvement may be only gradual and over a longer period during which novel criminal justice approaches are first devised, and then formally tested, to produce persuasive evidence about what does work at local or national levels.

2.2. Definition of indicators

PIs appropriate for achieving PM's primary objectives must be defined. Genuine consultation about the practicalities of implementing PIs is essential, and should be documented (see Section 2.5).

(a) Indicators should be directly relevant to PM's primary objective, or be an obviously adequate proxy measure.
(b) Definitions need to be precise but practicable.
(c) Survey-based indicators, such as of patient satisfaction or student placement after leaving university, should use a shared methodology and common questions between institutions.
(d) Indicators and definitions should be consistent over time. If, exceptionally, changes need to be made, these should be documented in a PM protocol (see Section 2.5), and an impact assessment made. Otherwise, a new PM base-line needs to be set.
(e) Indicators and definitions should obviate, rather than create, perverse behaviours.
(f) Indicators should be straightforward to interpret, avoiding ambiguity about whether the performance being monitored has improved or deteriorated. Ambiguity arises if there is major unaccounted-for disparity in case mix of individuals received into institutions, or if case mix (of pupils entered for examinations, patients selected for operation, or offenders given police cautions (House of Commons Public Administration Select Committee, 2003d)) drifts over time.
(g) Indicators that are not collected for the whole population should have sufficient coverage to ensure against misleading results, i.e. potential bias compared with measuring the target population should be small.

(h) Technical properties of the indicator should be adequate, e.g. with respect to sampling scheme, response rate in surveys, and precision (standard error of difference, say).
(i) Indicators should have the statistical potential to exhibit or identify change within the intended timescale of PM (see Section 2.3).
(j) Indicators should be produced with appropriate frequency, disaggregation or adjustment for context, and timeliness to support performance management.
(k) Indicators should conform to international standards if these exist, e.g. the UK's new consistent definition of drug-related deaths (Jackson, 2002).
(l) Indicators should not impose an undue burden—in terms of cost, personnel or intrusion—on those providing the information.
(m) Measurement costs should be commensurate with PM's likely information gain.
(n) PM protocol (see Section 2.5) should detail an analysis plan per PI which is commensurate with the cost, quality and importance of the acquired data.

2.2.1. Example (perversity)

A key PI for prisons is the number of 'serious' assaults on prisoners. Defining severity as 'proven prisoner-on-prisoner assault' risks the perversity (House of Commons Public Administration Select Committee, 2002a) of failing to investigate assaults on prisoners rather than trying genuinely to improve prisoner safety.

2.2.2. Example (definitions, drift and likely information gain)

Ambulance services differ in the methods used to classify emergency calls as life threatening and in when they start the clock for working out whether an ambulance arrives within 8 minutes to a life-threatening emergency (start time = call made, call answered, classification time, dispatch of ambulance). Ambulance services appear to have made perverse changes in classification methods or start times to manipulate the indicator of their performance without genuine enhancement (House of Commons Public Administration Select Committee, 2002b). Research on combining classification rules with local journey times to optimize survivorship may be a more cost-effective, and directly useful, solution than misjudged PM.

2.3. How to set targets

Targets are often linked to PM schemes. If performance targets are to be set, they need to have a sound basis and to take account of prior (and emerging) knowledge about essential variation. It is clearly appropriate, therefore, to deal with targets, when set, in the design of a PM scheme.

Indicators should have the statistical potential (known as 'power') to exhibit, or identify, change within the intended timescale of PM. Technically, this requires making a reasoned, prior assessment of how much improvement it is plausible to achieve within the PM timescale by taking account of research evidence, local initiatives, organizational culture, and the available or budgeted-for new resources.

Next, statistical power needs to be worked out in respect of this rational target. Aspirational targets have a distinctive role, but one which is largely irrelevant in the design of a PM scheme; motivational targets which are not rational may demoralize. As a general rule, modest changes need large studies for their identification. Setting up an indicator to detect a hoped-for 50% reduction in reoffending, when only a 5% reduction is realistic, would typically lead to a 100-fold underestimation of the number of offenders to be followed up. In practical terms, about four times as many patients, or pupils or offenders need to be surveyed, or tested or followed

up if the effect size of interest halves, and 100 times more if the effect size of interest is reduced by a factor of 10.

Equally, equivalence between two policies, which itself requires definition, cannot be claimed if a PI does not have the statistical power to identify even moderate effects (favourable or unfavourable), had they occurred as a consequence of changing policy. This size consideration applies particularly to PIs which are based on sample surveys (Gore and Drugs Survey Investigators' Consortium, 1999).

2.3.1. Example (power)

With 60 000 random mandatory drug tests (RMDTs) of prisoners per annum, a national prison service has excess statistical power to identify whether prisoners' opiate positivity decreases by 10% between calendar years from 5% to 4.5% of prisoners testing positive. Statistical power remains comfortably adequate (more than 80% power) for doing so with around 30 000 RMDTs per annum.

A different prison service with only 6000 RMDTs per annum but higher prevalence of opiate use has only 50–60% power to detect a 10% decrease between calendar years in prisoners' opiate positivity from 15% to 13.5%, but comfortable power between 2-year periods. Thus, statistical power depends both on number tested and base-line prevalence of opiate use. And so, the timescale over which power is sufficient may be different for different services.

Target setting should consider objectively whether a planned policy initiative or just routinely monitored performance is likely to have a dramatic, moderate or only modest effect on a chosen PI. Modelling in advance what may be achievable through policy initiatives can inform target setting, as can critical, systematic review of the scientific and grey literature on what has been achieved elsewhere.

2.3.2. Example (advance modelling)

In transplantation, proposed changes in the national rules for matching cadaveric donor kidneys to recipients are tested in advance by simulation; and, after implementation, actual *versus* projected impact on transplant survival and waiting times are compared (Armstrong and Johnson, 2000; Fuggle *et al*., 1999).

Equally important is to heed the dynamic environment within which a target is set or PI is measured. Substantial improvement in the provision of care for the elderly may be needed consistently to meet a target for the over 75 year olds, simply because of the UK's aging population. In education, there has been debate over 'inflation' of examination grades. More generally, if conditions have changed significantly within a service whose performance is being monitored, this needs to be taken into account at analysis, and targets also revised as necessary. There should, of course, be a documented audit trail of target or PI revisions and the reasons for them (see Section 2.5) which is open to independent scrutiny.

2.4. How not to set targets

Her Majesty's Treasury *et al*. (2001) coined SMART targets as specific, measurable, achievable, relevant and timed. Several approaches to target setting are so statistically problematic that we highlight them here.

First, it is unsmart to ignore uncertainty. A surgical team's 1-year mortality after cardiac artery bypass surgery should *not* be judged on whether 'number of deaths within 1 year/number

of cardiac artery bypass operations performed' (D/n) is less than target. Rather, judge the team's performance on whether an appropriate estimation interval for their 1-year mortality after cardiac artery bypass surgery includes the target.

Second, and for similar reasons, it is statistically unsmart to sharpen targets progressively by requiring that next year's performance is better than 'the better of current target and current performance'. Despite consistent performance, failure is inevitable. Moreover, uncertainty in 'current performance' has been conveniently, but wrongly, ignored by giving it target status!

Third, it is unwise to cascade targets by imposing the same target—for example, that at least 75% of pupils achieve a literacy standard—on a class of 30 as nationally, where class size is effectively 600 000 and precision of estimation is dramatically greater. For a class teacher to be 99% confident that at least 75% out of 30 pupils actually achieve the literacy target, the expected pass rate for his or her class would have to be close to 90%.

Fourth, it is usually inept to set an extreme value target, such as 'no patient shall wait in accident and emergency for more than 4 hours' because, as soon as one patient waits in accident and emergency for more than 4 hours, the target is foregone, and thereafter irrelevant. Typically, avoiding extremes consumes disproportionate resources. With similar intent, a more cost-efficient and continuously relevant target would be '95% of patients wait in accident and emergency for under 4 hours'. There are, of course, cost-justified important exceptions, such as prison services' target of zero escapes of category A (high security or dangerous) prisoners.

Fifth, it is unsmart to base waiting time targets on time cut-offs (under 4 hours or under 2 weeks) unless data on the distribution of waiting times are also collected because insights, such as crowding of waiting times at the cut-off or approximations to nearest 15 minutes, can be discovered from the distribution of actual waiting times, as in the Bristol Inquiry (Spiegelhalter *et al.*, 2002).

Sixth, it is unsound for target setting to ignore well-understood essential variation that is familiar to practitioners, such as young offenders' much lower rates of inside use of opiates compared with adult prisoners.

Seventh, it is unsmart to specify a national target in terms of statistical significance rather than in operational terms because, with a sufficiently large denominator, such as from complete enumeration of 600 000 school-children's test results, very small differences of no practical importance achieve statistical significance.

Finally, we remark that ignorance about the statistical properties of a new PI diminishes substantially after the first round of analysis, at which major sources of variation are likely to be identified (or confirmed). Quite properly, target setting may be delayed until the statistical properties of a new PI are sufficiently understood. Therefore, eighth, it lacks rigour for target setting to fail to document the level, precision and base-line period from which PI change is to be measured, as in 'By 2004, reduce by 20% the level of drug-related deaths from X (precision x) in 2000'.

2.5. Performance monitoring protocol

In principle, getting a PM protocol right is no more, or less, onerous than applying scientific method to any other field. The scientific discipline of writing a PM protocol from consultations and definition of PIs through piloting and data collection to auditing and analysis plans contributes to quality assurance of a PM process.

As in any other well-conducted study, the purpose of a PM protocol is to document not only the decisions about design which have been taken but also the underlying reasons for those decisions. In particular, besides objectives and definition of PIs, the PM protocol will record

the sorts of perverse behaviour that consultation with practitioners had given forewarning of, and the steps taken to obviate perversity. Survey methods and materials, target setting in the light of prior knowledge, and statistical power in respect of realistic targets together with the results of any necessary pilot studies will also be found in the PM protocol. How the PM data will be checked and audited, available information about context and case mix, the planned analyses and dissemination rules are other essential features of a PM protocol, which should also identify the PM designer and analyst to whom queries may be referred. Both for ease of reference and to enhance practitioners' and public understanding, analytical knowledge about the statistical performance of a proposed PI monitoring scheme under plausible assumptions about components of variation should be summarized in the PM protocol. Consideration may be given to how PM's cost-effectiveness will be assessed. Last, but not least, the PM protocol should also detail how ethical, legal and confidentiality issues have been resolved; see Section 7.

If the PM protocol reveals a dearth of prior knowledge about essential variation, as may be inevitable in respect of a newly defined PI, the protocol may specify a staged implementation of PM to allow essential variation to be well understood. The PM protocol also provides an audit trail of revisions to the PM process: when and why they occurred, as well as impact assessment. Revisions in the light of pilot studies should be anticipated in the overall timetable for PM implementation.

Failure to design, and audit properly, a robust PM protocol is false economy because to buy cheap methodology is to buy dear in the longer term (Gore and Drugs Survey Investigators' Consortium, 1999) if subsequent audit or independent critique discovers problems with performance data which have been influential in public policy debate (Audit Commission, 2003b). The serious indirect cost is loss of public and practitioners' confidence in the design and analysis of PM.

Focusing on PIs rather than the service can result in a type of statistical gaming whereby, instead of improvement in existing services, PM leads to service drift so that *either* individuals are excluded from receiving the service whose attendance would be likely to compromise PIs *or* an institution's range of services is limited in future to those associated with high past performance on measured indicators. Both described changes improve institutional rating without improving performance on a like-with-like basis. PM schemes should be so designed that statistical gaming is detectable.

PM can be used as the backdrop for intelligent sampling of institutions for follow-up inspection. The role of inspection (Fairweather (2002); see also `www.chi.nhs.uk`) is as follows: to discover, or generate testable hypotheses about, the reasons which underlie heterogeneous performances; to provide extra contextualization outside the PM scheme; and to disseminate good practice. PM protocol should therefore describe whether, and how, data on PIs will be used to sample institutions for follow-up inspection. For example, more institutions may be selected at random than are sampled from those

(a) apparently underperforming and
(b) apparently well performing.

And, maximally to preserve a level playing field at inspections, both inspectors and the institutions themselves may be blind to this apparent labelling.

Dissemination rules are an essential part of any PM protocol because they assure properly informed consent, the right of public oversight and consistency with National Statistics guidance (Office for National Statistics, 2002) in matters such as data checking, analysis feed-back to institutions, follow-up inspections, timing of publications, and the labelling or naming of institutions or individuals. In particular, the PM protocol can be expected to specify who holds the

key to performance identities, how institutions will be identified within the statistical centre to preserve confidentiality and objectivity at analysis, and the criteria to be met before institutional or individuals' identities will be revealed to third parties, and who these third parties are.

3. Analysis of performance monitoring data

3.1. General principles

The key principles for the analysis of PM data are no different from those for other types of data. The analysis should be sharply focused on the appropriate objectives and should be as simple as possible, subject to avoiding misleading oversimplification. We return in Section 4 to the presentation of results of the analysis but that too should be tuned both to the objectives and also to the potential users.

Technical issues in the analysis of PM data are part of the role of the statistician and should be treated as such if public confidence in the objectivity of the performance measures is to be maintained. In particular, it is essential that the compilation, presentation and statistical interpretation of performance measures should be seen to be impartial.

3.2. Data collection

The precise form of the data collection must be specified for each indicator of performance. This will include the choice of data sources and whether based on a sample or a census and, if the former, the sampling design. Precise definitions are required also of variables that allow adjustment for context as discussed later. The frequency and predictability of measurement (unannounced or announced) must also be specified in the light of the issues discussed later.

3.3. Data quality

A small amount of defective data may be quite misleading. A large amount of defective data may be extremely misleading. There is no security in quantity over quality.

Thus, key preliminaries to any kind of interpretation are checks of data quality. These should often be at two broad levels. One is to confirm, probably by some form of sample audit, that individual indicators by and large measure what they purport to measure and have not been corrupted by unanticipated perverse behaviours. The other is to isolate suspect individual values that are incorrectly entered into a database or which are subject to misunderstanding and which may seriously distort final conclusions.

3.3.1. Example

Some values may be detectable as logically impossible, e.g. percentages that should add to 100 do not or essentially female operations, such as hysterectomy, entered as referring to males. Missing day of month may be entered differently between institutions, e.g. as 00 in some but as 01 elsewhere, inviting error or illogical date sequences (death prior to treatment, say), if the second is taken at face value. More generally, there is a well-developed set of statistical principles and procedures aimed at detecting and dealing with outlier or aberrant observations.

We assume in the following discussion that such checks have been effective.

3.4. Importance of variability

The next important point is that, in most cases, analyses need to examine not just overall average values of PIs but to look at variability. This is for two rather different reasons. One is that the

variability is an intrinsic part of the real world and often of interest and importance in its own right. The other is that such variability is one of the determinants of uncertainty in the primary conclusions.

3.4.1. Example

Suppose that cancer mortality rates are to be compared as between patients in specialized cancer hospitals and cancer patients in general hospitals. An overall comparison of two rates, one for each type of hospital, might conceal substantial variation between hospitals of the same type. This variation might well be of intrinsic interest. At the same time the variation would indicate the level of confidence that could be attached to a general conclusion about the merit of specialized hospitals.

Note that the use of census data does not preclude the need to consider variability.

Analysis should also distinguish between sources of variation that an individual practitioner or organization has control over in the short term and those that are modifiable only externally or in the longer term.

3.5. Unit of study

In terms of detailed analysis and interpretation, a further key choice concerns the unit of study, which should at least be consistent with PM's designed objective.

3.5.1. Example

In analysing data on reading performance, the unit of study might be an age group within a school, a school, or all schools in a local authority area. All are potentially relevant, but for different purposes.

Analysis may need to take account of organizational or some other hierarchy that is known, or discovered, to determine essential variation. For example, much of the variation in school pupils' performance takes place at the classroom level, which cannot even be captured at analysis unless classes within a school were distinguished when the PM data were collected.

3.6. Frequency of analysis in repetitive situations

In some contexts PM is done relatively infrequently and analysis will concentrate on the current data recently collected with perhaps some brief comparison with earlier data. In other situations PM is quite frequent, perhaps almost continuous. Here techniques analogous to the industrial control chart may be suitable. Each new value is plotted to show its relation with previous values and usually there will be warning lines to indicate unsatisfactory performance, in particular deteriorating performance or failure to reach targets, in the industrial case specification limits. In many industrial contexts there has, however, over recent years been a shift from the relatively passive role of detecting failure (naming and shaming, in a sense) to a much more positive role of encouragement of constant improvement by evolutionary operation, as suggested many years ago by G. E. P. Box (Box and Draper, 1969) when at ICI Ltd. This involves repeated, organized modification of procedures and calls for a very slightly more elaborate analysis. While some of the more recent ideas of this kind are presented in a very hyped-up fashion, it is possible that much of the thinking, especially the emphasis on encouragement, could be applied in very different contexts.

3.7. Adjusting for context to achieve comparability

Analysis beyond simple tabulations and graphical summaries is unavoidable when it comes to adjustment for context, also known as risk or prior status or case mix adjustment. This is often a critical part of analysis. Sensitivity to assumptions needs examination and the process of adjustment must be transparent to be convincing to the user and to avoid fruitless controversy. Also, there needs to be recognition that case mix may alter over time, because selection criteria change or because risk is age related (and therefore liable to be increasing because of demographic aging).

3.7.1. *Example*

If crude mortality rates were computed annually for individual cardiac surgeons, a practice we emphatically discourage, it would be absurd to compare the rate for a surgeon specializing in immediately life-threatening cases with one dealing with much more routine procedures. Even in less extreme cases, however, and even if the unit of study is a hospital department or even a whole hospital, comparability of different rates would be far from clear. Context adjustment attempts to ensure that comparisons can be directly interpreted by, so far as possible, comparing like with like (Bridgewater *et al*., 2003).

There are broadly two ways in which context adjustment can be attempted—by simple stratification or by statistical techniques for regression or multilevel modelling which in essence carry out a form of stratification simultaneously on several, or indeed many, features using an empirically estimated equation to represent the adjustment effects.

3.7.2. *Example*

To assess the general academic performance of, say, 15 year olds a number of PIs will be available. Suppose the unit of study is the school and all that is known about performance at entry to the school is the classification of each pupil as of high, medium or low attainment. Then to compare schools by a single measure we, in effect, compare the low attainers across schools, the medium attainers and so on and, if appropriate, then combine the answers into an overall measure. Note, however, that the detailed analysis allows also the possibility of assessing effectiveness with different types of pupil.

A central issue is the choice of context variables on which to base the adjustment. They should represent external features outside the control, judgment or influence of the organization under assessment. This can be difficult to determine without local knowledge, or audit.

3.7.3. *Example*

Some transplant units may be more likely than others to offer a non-favourably matched, local kidney to a patient who has been waiting for less than 2 years but who, because of tissue type, has only a modest chance of being offered a very well-matched kidney by UK Transplant in the next 5 years. In such a case, the possible adjustment factor, closeness of tissue matching, is not entirely outside the control of the units being monitored.

The uncertainty associated with any method of risk adjustment needs to be kept in mind and perhaps formally incorporated into monitoring procedures. In particular, the incompleteness of any method of adjustment must be recognized. Therefore, when management decisions are

based on an analysis of risk-adjusted performance, allowance must be made for relevant factors either not captured in the database or which may not have been optimally dealt with by the chosen procedure for context adjustment.

3.7.4. Example

The main reason for not publishing mortality rates for individual surgeons is the difficulty of accurate and relevant assessment of the patient mix that each surgeon encounters and of the seriously bad consequences that could follow from misinterpretation; see Section 7.

3.8. Multiple indicators

As discussed previously there will often be available for analysis several or indeed many PIs. While for some purposes, e.g. resource allocation, it may be necessary to amalgamate these into a single summary figure, this amalgamation should be resisted as far as possible and, if essential, postponed to a final stage (Audit Commission (2002); see also `www.councilperformance.gov.uk`). Performance assessment may be severely distorted by amalgamation. Also, value judgments are implicit in any amalgamation, but these may differ legitimately between stakeholders.

On the other hand, if a large number of PIs is collected, some summarization is essential in analysis. For this we recommend, as illustrated in the example of kidney transplant data discussed in Section 2, forming a limited number of dimensions describing distinct features of the system, and calculating summary scores for each dimension, corrected for context where appropriate. The number of dimensions must be specific to the particular situation. Then, whether the objective is a comparison across time, a comparison across similar institutions or the formation of a league table, results for the different dimensions are shown in a directly comparable form. This has a number of advantages. It precludes overemphasis and distortion often inherent in single measures and it also gives some indication of intrinsic variability, totally hidden in, for example, a conventional league table based on a single summary measure across dimensions.

3.8.1. Example

For spider-webs to illustrate aspects of police performance, and otherwise known as star plots and stardinates, see Chambers *et al*. (1983) and Audit Commission (2003c) (see also `www.policereform.co.uk/performance_monitors/monitors_intro.html`).

3.9. Indicators measured over time

Some PIs are calculated at relatively frequent intervals, e.g. examination successes every year. In such cases it is important for interpretation and subsequent decision-making to study the sequence of values over a suitable time range. How long a range must depend on the context, in particular how long the system has been operating in broadly its present form, but 5–10 time periods will often be adequate. There is a clear tension here between PM's aim of identifying contemporary competence among practitioners or institutions and statistical potential to do so. Very particularly the practice of concentrating on a comparison with the most recent value, this year's results compared with last year's, may be very misleading for reasons including regression to the mean. Even if the objective is to assess the impact of a major recent innovation, comparison with just one past value on its own, rather than with a recent average level or trend, will often be unwise. In other cases it will tend to encourage the overinterpretation of very minor changes which may be seen as of no lasting importance in the light of the general variability involved.

Summary measures of trend are sometimes advocated for the analysis of time series data. However, even with longer time series, it can be very difficult to estimate trends very precisely. Thus, in this situation as in others, care must be taken to represent adequately the uncertainty associated with any summary measure of data acquired over time, including the sensitivity of the summary measure to inclusion or exclusion of the more historical data.

3.10. Summary on analysis

In general terms, the effort spent on analysis and interpretation of data should be commensurate with the effort spent in collecting the data and with the importance of decisions which will explicitly or implicitly be based on the data. Even when the data are collected primarily for individual monitoring and encouragement, simple and fair analysis and presentation remain important.

There is a large variety of special statistical techniques that may be used for simple summarization, for context adjustment and for the assessment of the uncertainties involved; see Goldstein and Spiegelhalter (1996). This is not the place to go into the details of such techniques but it will be clear that we favour utmost simplicity consistent with efficiency, clarity and security of interpretation. Even when efficiency or security of interpretation requires that complex techniques are used, this does not preclude either intelligible lay description of the underlying principles or a widely accessible presentation of actual results.

Sampling approaches may add slightly to the complexity of analysis but their potential cost-effectiveness and safeguarding against perverse behaviours commend them as sometimes both more accurate and less costly than complete enumeration.

4. Presentation of performance indicators

4.1. The purpose of presentation

When a new PI is introduced, initial analyses should be designed to provide insights into its characteristics, particularly essential components of variation, which should inform decisions concerning its routine presentation. Performance management, of some type, may be the ultimate goal to which routine analysis and presentation of particular PI data contribute. Inspection, decision costs and consideration of 'value for money' may be needed to decide what to do. For the purposes of performance management, direct, and possibly naïve, use of PI information may be ineffective, or even counter-productive.

4.2. General principles

The principles of presentation are not really different for performance data from other sorts of data. There must be a strong focus on the real objectives and the simplest mode of presentation that avoids being misleading.

It is virtually always necessary that some direct or indirect indication of variability is given. Simplicity does not mean discarding measures of uncertainty either in tables or figures. Insistence on single numbers as answers to complex questions is to be resisted. The final synthesis of measures from several dimensions into a single index, if necessary for some final decision, should be postponed to as late as possible in a decision process and the sensitivity of decisions to the relative weighting of the different dimensions examined.

Where the conclusions are for immediate action or discussion graphical methods will typically be best; where further analysis and comparison may be involved a tabular mode may well be preferred. Mode of presentation should be considered when the collection of PM data

is designed and in planning the analysis so that the primary method of presentation can be prespecified in the PM protocol. Public availability of any software to be used in monitoring may also be helpful in developing confidence in the methods used.

4.3. League tables of performance, or of change in performance

The limitations of published league tables which explicitly rank individuals or organizational units on the basis of PI data are well known. For example, a strong case for the imprecision of the ranking is made by Goldstein and Spiegelhalter (1996). League tables are frequently produced on a yearly basis when a better basis for ranking or estimation may be achievable through use of a more extended period of observation. However, in all cases, the uncertainty of ranking should be indicated through the use of plausible ranges of rank for each institution, or some comparable approach. Even if only broad banding of ranked individuals is to be presented, the band definitions must incorporate adequate allowance for uncertainty. League tables without demonstration of ranking uncertainty should be avoided and, even with uncertainty incorporated, such tables should be used with considerable caution.

4.3.1. Example

For a league table with broad confidence intervals about ranks, see Fig. 1.

Classification of institutions into a few categories, e.g. 'star' banding of hospitals, may be used as an alternative to complete ranking. Whether based on a single PI or a balanced score-card of them, the choice of 'boundaries' for each PI in defining bands itself introduces controversy. Banding does not circumvent the need to represent the uncertainty of an institution's banding.

4.4. Extremes

Another perspective on ranking arises by consideration of 'extreme' units. If this is the primary focus of a presentation of PIs, then alternative methods for the presentation of uncertainty may be preferable. For example, the relevant question may not be what 'absolute' rank should an institution have, but rather whether in the set of institutions examined the lowest ranked ones are, in fact, worse than the worst are expected to be. A technical answer will usually involve distributional assumptions that require empirical support to ensure robust conclusions. However, the principle that being ranked lowest on this occasion does not immediately equate with genuinely inferior performance should be widely recognized, and reflected in the method of presentation.

4.4.1. Example

Spiegelhalter (2002) presented a 'funnel plot' of emergency readmission rates for stroke patients plotted against the number of stroke patients treated per year with exact binomial control limits which reflected the expected dependence of variability in readmission rate on sample size. Divergent hospitals stand out but ranking of the remainder is not implied; see Fig. 2.

4.5. Time series

The presentation of changes in PIs over time, particularly for the purpose of assessment of an intervention, should reflect the analysis principle that the comparison of a current value with just one past value can be misleading. While some allowance can be made for this through

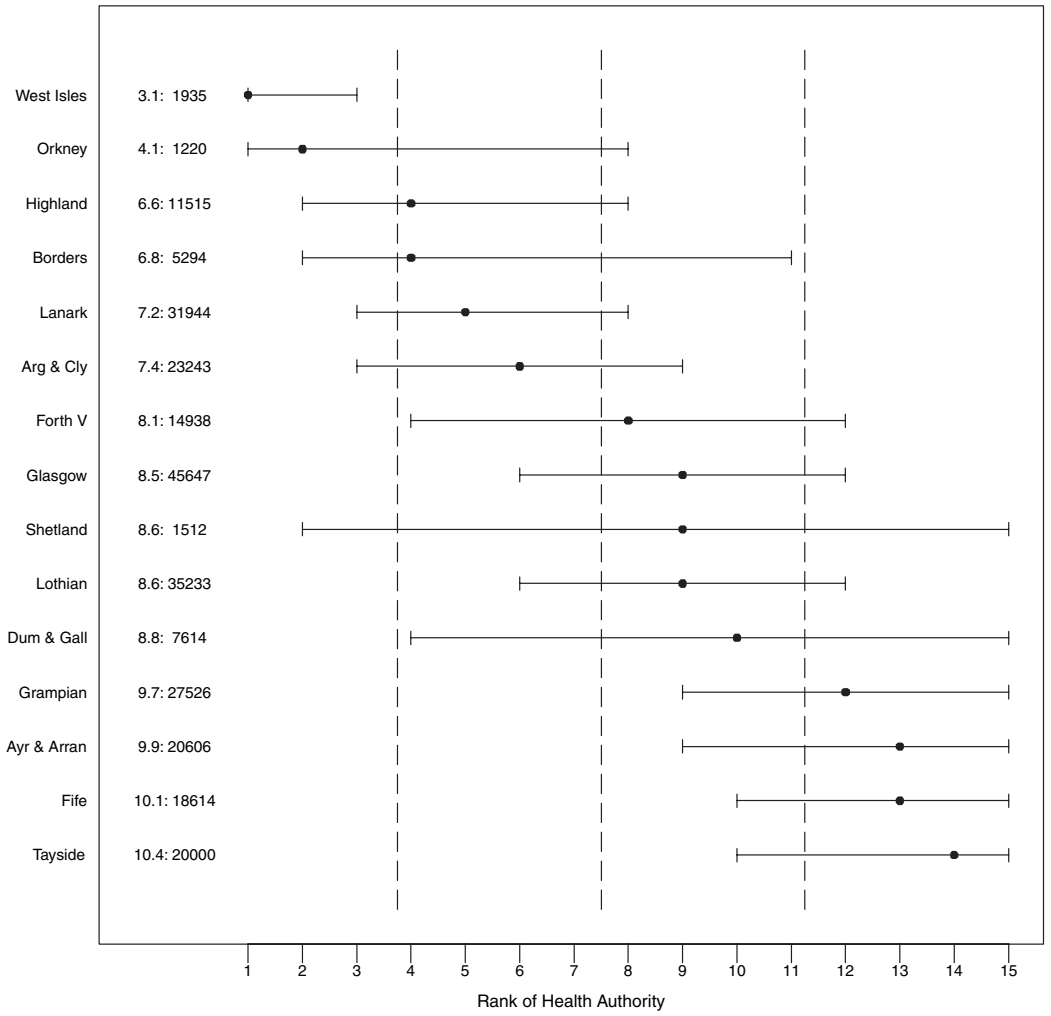

Fig. 1. Median and 95% intervals for ranks of Scottish health authorities with respect to teenage conception rates 1990–1992: rates and relevant populations are shown for each health authority (adapted from Goldstein and Spiegelhalter (1996))

statistical modelling, it can often be more simply accomplished through presentation of longer time periods of observation.

Plotting PI data as a series of points in time serves several purposes. First, it shows the natural variability of the PI. Secondly, by use of warning lines or control limits, it identifies outlying values that are likely to be unsatisfactory. Thirdly, it allows the precision of overall summary measures of practitioner or institutional performance to be assessed for comparison with external criteria, including targets.

For other purposes, in particular for the presentation of periodic data on multiple units, the use of interval estimates may be more natural. As discussed in Section 3.9, over-interpretation of apparent trends can be avoided by adequate consideration of the uncertainty associated with summary measures.

When risk-adjusted information is presented, the dependence of uncertainty measures on the adequacy of the risk adjustment should be considered. There is a strong case for presentation

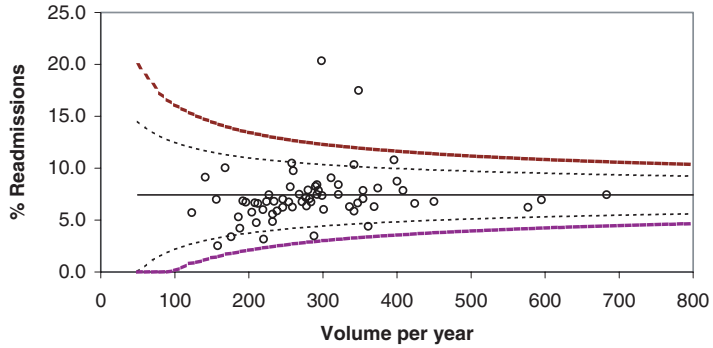

Fig. 2. 'Funnel plot' of emergency readmission rates following treatment for a stroke in large acute or multi-service hospitals in England and Wales in 2000–2001 (reproduced from Spiegelhalter (2002) with permission from the BMJ Publishing Group)

also of the data used for risk adjustment, as changes in these over time may be induced by the adjustment strategy adopted for PM.

4.6. Public release versus private feed-back

The decision whether PI data should be publicly released depends on a variety of factors. Public release based on the presumed value of a 'name-and-shame' approach is called into question by careful analysis following New York State's disclosure of patient outcomes at the level of individual, named cardiac surgeons (Dranove *et al.*, 2002). Analysis revealed surgeons' subsequent risk averseness and deterioration in patient outcomes. More research is needed on dissemination rules. 'Naming' is not a prerequisite for public accountability and may have disbenefits besides its apparent attractiveness in promoting public choice.

In addition, as mentioned earlier, PM may be viewed as having the positive role of encouragement of constant improvement. This role is not facilitated by frequent, named release of PI information or if vital clinical priorities are distorted because managers fear to fail so-called 'P45 targets' (House of Commons Public Administration Select Committee, 2003d; Hawkes, 2003).

In clinical trials, monitoring of results is publicly accountable but important information on treatment efficacy is most efficiently acquired by maintaining confidentiality about interim monitoring results unless firm conclusions can be drawn on efficacy, or there is a safety concern. Thus, while public scrutiny of PIs is an important principle, its implementation must be informed by careful consideration of the impact of strategies for public release (Yang *et al.*, 1999). The mechanism for public release should be specified in the PM protocol when any monitoring procedure is implemented.

4.7. Requirements for different audiences

Presentation of results should be understandable by the intended users. A presentation of PI data that is intended for public release should have intuitive appeal. Indications of variability such as control limits or confidence intervals may involve some statistical complexity in their determination but are usually adequately represented in a simple fashion. As discussed earlier, if PI data are being presented for many organizational units, a method that does not lend itself to ranking is preferred.

Feed-back to individuals or institutions more familiar with statistical concepts may allow somewhat more complexity in presentation. This can be particularly important when risk adjust-

ment is necessary or multiple PI measures are relevant. However, complexity or difficulty of public understanding should never be an excuse for insufficiency of analysis.

4.7.1. Example

For monitoring deaths following surgery by a single surgeon, risk adjustment via a likelihood ratio argument is optimal for cumulative sum charts (Steiner *et al*., 1999). Nevertheless, the intuitive simplicity of plotting cumulative observed minus expected deaths has wider appeal. Presentation of both is the obvious solution. If deaths are so rare that their monitoring risks missing actual changes in performance from lack of statistical power, a second PI may have to be introduced. In neonatal arterial switch operations, the occurrence of a so-called 'near miss' during an operation is used as a surrogate outcome; and formal methods for simultaneous monitoring of deaths and near misses have been developed.

4.8. Sensitivity and robustness

Some presentation of the sensitivity of PI analysis to key assumptions, including the choice of data for adjustment, is desirable. In particular, the robustness of banding is a clear requirement if decisions on performance management are informed by, or even solely reliant on, the analysis.

5. The impact of performance monitoring on the public services

Benefits of PM include new investment in data capture, common methods of measurement across institutions, and availability of between-institution comparisons when previously comparison could be made, if at all, only within institution and against *either* performance thresholds set by professional consensus *or* national targets.

There is also evidence that PM can secure specific changes in the measured aspects of public services, as intended. For example, reductions in monitored waiting times in the NHS are almost certainly due in part to the high profile given to waiting in the PM regime, and this is also at least partly the case with the increase in educational key stage test scores over time as schools attempted to meet externally imposed targets. PM may also change the behaviour of citizens in choosing health care or which school catchment area to live in. School test scores are even used by estate agents to promote house sales! PIs also feature in political debate and influence behaviour through, for example, local or national elections. Furthermore, managers and practitioners may find that PM gives them a clear preference in terms of policy priorities, whether or not they agree with those priorities.

Direct costs of PM include the cost of PM-specific data acquisition and of auditing the local implementation of PM protocols (Bird *et al*., 2002). Acquisition costs may be considerable, as with universities' participation in research assessment exercises (Berkshire, 2001). Survey-based indicators incur both sampling and survey costs. There is then the cost of data checking and analysis and of follow-up inspections. The cost of ameliorating actions in response to PM signals is a further consideration.

There is, of course, debate about the impact of PM beyond direct costs. For example, there is concern that unmeasured aspects of the NHS and education suffer. In addition, PM can also give rise to other unintended consequences (Smith, 1995) that may serve to thwart the intentions of policy makers, and work to the detriment of the public services. It can lead to manipulation of data, gaming or fraud by service providers. It can inhibit new approaches to service delivery. In complex systems, partial PM can lead to suboptimal global solutions. And the use of inappropriate PM may demoralize and undermine those charged with delivering public services.

Many of these negative consequences occur because a strong feature of the justifications for the use of PM is the assumption that the *process* of measurement does not influence the behaviour of individuals and institutions involved. This assumption, however, is questionable and there is now evidence, especially from health and education, that 'high stakes' performance assessment does indeed affect behaviour, and such side-effects are often counter-productive. Thus, among cardiac surgeons in New York whose individual unadjusted patient death-rates have been published regularly, there has been a tendency to avoid taking on high risk cases with a subsequent increase in mortality of Medicare patients at risk for cardiac surgery (Dranove *et al*., 2002). In the State of Texas a programme of rewarding schools and teachers based upon published student test scores has been shown to have produced dubious results, despite apparently very rapid increases in test scores overall as a result of 'teaching to the test' (Klein *et al*., 2000).

Behaviour change is a factor because no PM scheme can be viewed in isolation from the incentives—designed or accidental—that exist alongside it. Designed incentives often take the form of targets, and a set of consequences associated with performance. If the assessment of management functions in the NHS depends centrally on whether explicit waiting time targets are secured, then this can affect such things as patient handling strategies among health care professionals not directly involved but whose activities contribute to the targets. Public disclosure of police force performance may not be associated with any formal set of incentives, but—given the high media profile of the performance data—it would be surprising if police forces did not make some changes in response to the data.

As emphasized in Section 2, the precise design of any formal target is important. For example, the target of requiring no breaches in a waiting time guarantee for NHS patients involves a different resource allocation from a target of requiring that 90% of patients satisfy a more stringent waiting time criterion. In education, the publication of an examination PI consisting of the percentage of students achieving five or more 'good' examination passes at age 16 will affect the policy of schools towards potential 'borderline' candidates. On the other hand, publication of an indicator based upon an overall average points score will result in different emphasis being given by schools dependent on the actual weighting function chosen. There should be an attempt to anticipate how targets are likely to, and to monitor how they do, shape behaviours.

In the same vein, when a suite of targets is specified (as for schools), responses may differ considerably depending on whether organizations are assessed on each measure individually, or on a composite of all targets. Construction of the composite requires value judgments about relative importance of each target, but leaves organizations free to decide along which dimension to seek improvement. Nevertheless, the choice of any particular composite measure for summarization will have behavioural implications that should be considered and monitored.

PM risks overemphasis on the easily measurable and has the potential to conflict with the priorities and values held by professionals and/or citizens. For example, the emphasis on NHS hospital waiting times contradicts the pre-eminence given to clinical prioritization in the training of most health care professionals. However, it may be precisely because policy makers wish to temper professional values with other concerns that they feel a need for PM. This was the case in education where successive Governments had expressed concerns that teachers' values do not correspond sufficiently closely to a view of education that emphasizes its contribution to economic prosperity.

There are costs to the public if PM fails to identify under-performing units so that no remedial action is taken. There are real but less-well-recognized costs of falsely identifying a unit as underperforming when in fact there is no significant difference between it and others judged as

'adequately performing'. There are effects on staff, staff morale and recruitment which in turn impact negatively on the service provided. Extensive management and organizational changes, if triggered by false identification, have disruption costs that will not be justified by improved performance in the longer term.

Care should be taken not to undermine professional values unnecessarily and to demoralize and antagonize staff whom the public services rely on for delivery. This is an aspect of the need for a balance to be struck between short-term (and possibly only apparent) gains and long-term effects of any PM system.

In summary, when a new PM protocol is proposed, a broadly based assessment should be made of its likely costs and consequences.

6. Evaluating performance monitoring initiatives

Evaluation should be commensurate with risks and costs, not only of the PM initiative itself but also of the policies to be judged by a PM initiative (Smith, 1996). However, clear-cut deficiencies in a PM protocol need immediate, not evaluation-delayed, remedy. Alternative cost-efficient use of resources, such as for rigorously designed basic, clinical, educational or operational research, needs to be considered.

Four common difficulties in evaluating PM initiatives are when to start evaluation, constrained choice of evaluation design to conform with predetermined policy 'roll-out', confounding of a PM initiative with other policy initiatives so that their separate effects cannot be disentangled, and cost-effectiveness considerations. We highlight the role of experiments in PM (Propper and Wilson, 2003).

6.1. Difficulties of evaluation: how soon to start

If legislation is required to enable a particular form of PM, e.g. because it is intrusive (such as drugs testing of offenders) or costly (such as post-mortem bovine spongiform encephalopathy (BSE) testing of adult cattle), it may be difficult, or impossible, for implementation of a PI to be fully worked out in advance of its mandated start date. Consequently, data for the base-line year may be incomplete, and precision less; or targets may have been set prematurely, before sufficient knowledge about essential variation had been gleaned.

6.1.1. Example

To monitor the performance of BSE case detection by farmers and veterinarians, rapid post-mortem BSE testing of cattle came into effect from January 1st, 2001, with all cattle in the European Union slaughtered for human consumption at 30+ months of age subject to rapid BSE testing by one of three approved post-mortem tests. It took time to implement testing at abattoirs so that not until the second half of 2001 were BSE test results reported comparably in most member states (Bird, 2003). Without the need for targets, huge insight was gained by this PM initiative: BSE test positives at abattoirs are at least as numerous as clinically detected BSE cases and back-calculation of the UK's estimated past BSE epidemic has more than doubled as a consequence of active BSE surveillance.

6.2. Difficulties of evaluation: predetermined policy roll-out as design constraint

Wider consideration of, and education on, feasible evaluation designs to secure unbiased information about the impact of a new PI-monitored policy are needed. This applies especially when implementing the policy (its so-called roll-out) is subject to budgetary and other, including political, constraints.

Policy (a literacy initiative, say) and PIs (school league tables, say) often conflate in public debate. 'Pilot, then roll-out' can mean: pilot policy, then roll-out nationally both the policy and PI by which the policy's impact will be judged; *or* pilot PI, then roll-out nationally the policy which PI has been designed to monitor; *or* pilot policy and PI simultaneously before both are rolled out nationally. Because objectives are generally different when evaluating a PI and evaluating a policy, the impact of which will be monitored by change in PIs, separate evaluation designs are needed to meet these two distinct objectives. Conflation is unhelpful.

6.3. Difficulties of evaluation: confounding of policy effects

If PM is one of several initiatives introduced simultaneously or in rapid succession, all designed to achieve the same performance target, then PM's specific contribution is confounded.

Also, roll-out of a clutch of policies, and the associated funds for their implementation, is common, and often achieved via the same set of 'favoured' or purposely selected institutions, be they schools, hospitals, police units or prisons. More research is needed on efficient evaluation designs, which conform to overall or policy-specific funding and other constraints, yet allow unbiased policy-specific information to be obtained when a related series of PI-monitored policy initiatives is 'rolled out'.

There is ample justification for a methodology working group, as between National Statistics, the Treasury and Delivery Unit, to suggest feasible evaluation designs for Government departments to use when rolling out a series of PI-monitored policies, and for the Treasury to take cognizance of them in setting budget constraints.

6.4. Difficulties of evaluation: cost-effectiveness matters

Wider consideration needs to be given to how PM's cost-effectiveness should be appraised, and to cost-effectiveness in performance generally. Apparent underperformance may be due to under-resourcing. Excellence may have been achieved only at high cost—generally, or atypically at the time of performance assessment (House of Commons Public Administration Select Committee, 2003b) which we refer to as input gaming. Cost-efficient institutions may go unrecognized unless outcomes are related to input costs.

If a PM initiative is itself costly to implement in relation to the likely gain from it in terms of new insights or performance enhancement, then it should be discontinued for exceeding reasonable 'value-for-money' thresholds.

6.4.1. Example

The NHS can seldom afford licensed pharmaceutical drugs of proven efficacy whose cost-effectiveness is above £30000 per quality-adjusted life-year gained. Should £150000 expended on PM in the NHS therefore buy the equivalent of five full-quality life-years gained for patients, or have other comparable benefits?

6.5. The role of experiments, randomization and qualitative studies in performance monitoring

In addition to efficient designs for evaluating 'rolled-out' policies, a range of research questions in relation to PM could be resolved by experiments (Propper and Wilson, 2003), which ideally include random allocation (Smith, 1996). The list might include reporting format for PM (public anonymity of all institutions *versus* anonymity except by previously agreed criteria), choice of

follow-up actions to PM (action A *versus* action B), PM-based comparison of policies (institutions randomized to implement policy C *versus* policy D), prioritization of PIs (institutions randomized to prioritize set E {nationally prioritized five} *versus* institution-prioritized set F {also five in number} out of 20 PIs), or economic trial of PM *versus* alternative use, say for operational research, of equal resources (over 3 years) to achieve some 5-year objective.

Formal qualitative or other anonymity-assured study of how PM is perceived by practitioners, of behaviour changes that PM actually induces, and of altered allocation of resources, changed priorities or other reorganizations as a consequence of comparative information about institutional performance would ensure that the impact of PM was better and sooner understood. In particular, perceived or actual bullying could be allayed by responsive changes to the PM protocol.

7. Integrity, confidentiality and ethics of performance monitoring

Publicly accountable PIs need wider consensus than just from within Government, and also need independence safeguards for their design and monitoring, as is in principle achieved for National Statistics. Better public understanding both of uncertainty and of adjustment to compare like with like would help to ensure that PM-informed choices and actions are soundly evidence based. Confidentiality, ethics and properly informed consent should be explicit considerations in PM, as in other research which requires access to data on humans.

7.1. Integrity of performance indicators for public accountability

When PIs have been selected to assure public accountability for the Government's stewardship of the public services, then Parliament and the public need confidence that these statistical indicators are free from inappropriate political interference in their design and monitoring. This calls for all the safeguards that the Government has espoused in the creation of an independent National Statistical Service (Her Majesty's Government, 1998, 1999; Office for National Statistics, 2000, 2002; Statistics Commission, 2003b). In short, PIs for public accountability should be equivalent to National Statistics (Moore, 1991; Statistics Commission, 2003b). There is an unsatisfactory circularity about Government itself deciding on the PIs for public accountability. It risks the side-stepping of both wider consensus and the independence safeguards that are required of National Statistics.

A process is needed that engages experts and others, outside as well as inside Government and the public service itself, to build a more widely based view of the appropriate PIs and that resolves the inevitable tension between idealized objectives and what is statistically feasible.

Once PIs for public accountability have been determined, there are clear benefits to Government, institutions (e.g. schools, universities, hospitals, police forces or prisons) and the political process generally, if arrangements for their collection and publication are explicitly shielded from direct political influence.

7.2. Performance monitoring and statistical integrity

PM lacks statistical integrity if it fails to identify, and to design out, major corruptions of measured indicators, if its chosen PIs are not measured with sufficient precision to reveal whether targets have been met, if targets are set irrationally from the perspective of prior evidence, if its design is cost inefficient, or if analysis lacks objectivity or is superficial.

The drawing of wrong conclusions by faulty analysis or data collection and definition is, of course, to be lamented. But, designers of a PM procedure must do more than wring their hands:

when problems are uncovered, whether by audit (Audit Commission, 2003b) or independent scrutiny, modifications to the PM procedure must be speedily introduced.

The role of the statistician in PM is *both* strenuously to safeguard from misconceived reactions to uncertainty those who are monitored *and* to design an effective PM protocol which allows data to be properly collected, exceptional performance to be recognized and the reasons for it to be further investigated. On the basis of PM outcomes, the statistician can propose an efficient random sampling scheme by which institutions are selected informatively for announced or unannounced follow-up inspections.

7.3. Performance monitoring, confidentiality and ethics

Frequently, PM requires access to data about people (e.g. patients, school-children, prisoners or employees). Generally, research on humans needs third-party approval of its ethics and methodology, and is respectful both of confidentiality and the properly informed consent of participants. Exceptionally and for good reason, individual consent may be relaxed in the public interest, or for specific benefit. But, it is necessary to ensure that a research participant's identity cannot be deduced from published data, or analyses.

The corresponding problems connected with PM are major and have been little discussed. Human rights, data protection, freedom of information and, by contrast, legislation proposing access to personal data (Her Majesty's Government, 2003) for audit and inspection commissions need scrutiny. Moreover, while staff who deliver public services may well have agreed implicitly to the monitoring of their performance, it is far from clear that this extends to publication of analyses in a form in which individuals can be identified. In some contexts, this may be highly counter-productive in inducing a climate of fear or disaffection. In others, publication may be essential to achieve key objectives.

There are, however, legalistic requirements for the protection of confidentiality arising from the United Nations's 'Fundamental principles on official statistics' (United Nations, 1994) and the Amsterdam Treaty within the European Union (European Union, 1997).

Legal and human rights liabilities for serious, unwarranted costs to the reputation of institutions or individuals that may result from publicly disseminated, named league tables which do not properly convey uncertainty or from incorrect public labelling of an institution as 'failing' on the basis of a superficial analysis are relevant other considerations.

Principles which public dissemination of information about the performance of identifiable institutions or individuals should respect are the following: an institution's or individual's right to verify data; right of prior access to analysis plans, including for risk adjustment; required presentation of uncertainty; respect for multidimensionality of performance; consideration of cost-effectiveness of performance; avoidance of unwarranted harm such as by inspections prior to named dissemination; public right of access to robust PIs which inform the public's choice, or local and central funding, of the public services.

7.4. Public understanding of performance monitoring

The House of Commons Public Administration Select Committee (2003b) called for a more mature political debate about 'the measurement culture' and for a better understanding of targets as tools to improve performance. Central to that maturity is public, as well as political, understanding that there is inherent variability in all PIs, however well designed, which cannot be ignored. Even if a surgeon's ability is constant and the number and case mix of patients on whom she or he operates are identical this year and next, her or his actual number of operative successes need not be the same this year and next—owing to inherent variability and despite

constant ability. The Royal Statistical Society considers that education of the wider public, as well as policy makers, about the technical issues surrounding the use, including internationally (World Health Organization, 2002), publication and interpretation of PIs is very important. High priority should be given to sponsoring well-informed public debate, and to disseminating good practices by implementing them across Governments.

Acknowledgements

The Working Party is grateful to Fellows of the Royal Statistical Society and others who commented on our consultation report. We thank David Spiegelhalter and the BMJ Publishing Group for permission to reproduce the figures.

References

Armstrong, S. and Johnson, R. (2000) New National Kidney Allocation Scheme—first year monitoring report summary. *UKTSSA Users' Bull.*, **35**, 5.

Atkinson, T., Cantillon, B., Marlier, E. and Nolan, B. (2002) *Social Indicators: the EU and Social Inclusion*. Oxford: Oxford University Press.

Audit Commission (2002) *Comprehensive Performance Assessment: Scores and Analysis of Performance for Single Tier and County Councils in England*, p. 12. London: Audit Commission for Local Authorities and the National Health Service in England and Wales.

Audit Commission (2003a) *Memorandum PST 31 to the House of Commons Public Administration Select Committee*. Audit Commission, London.

Audit Commission (2003b) *Waiting List Accuracy: Assessing the Accuracy of Waiting List Information in NHS Hospitals in England*, p. 9. London: Audit Commission for Local Authorities and the National Health Service in England and Wales, London.

Audit Commission (2003c) *Waiting for Elective Admission: Review of National Findings*, p. 11. London: Audit Commission for Local Authorities and the National Health Service in England and Wales.

Aylin, P., Jarman, B. and Kelsey, T. (2003) What hospital mortality league tables tell you. *Br. Med. J.*, **326**, 1397–1398.

Berkshire, F. H. (2001) Heisenberg's uncertainty principle and quality assurance. In *Statistics, Science and Public Policy*, vol. 5 (eds A. M. Herzberg and I. Krupka), pp. 105–112. Kingston: Queen's University.

Best, N. and Day, S. (2004) Foreword to the papers on 'Performance monitoring and surveillance'. *J. R. Statist. Soc.* A, **167**, 447–448.

Bird, S. M. (2003) European Union's rapid BSE testing in adult cattle and sheep: implementation and results in 2001 and 2002. *Statist. Meth. Med. Res.*, **12**, 261–278.

Bird, S. M., Pearson, G. and Strang, J. (2002) Rationale and cost-efficiency compared for saliva or urine testing and behavioural inquiry among three offender populations: injectors in the community, arrestees and prisoners. *J. Cancer Epidem. Prevn*, **7**, 37–47.

Box, G. E. P. and Draper, N. R. (1969) *Evolutionary Operations*. Chichester: Wiley.

Bridgewater, B., Grayson, A. D., Jackson, M., Brooks, N., Grotte, G. J., Keenan, D. J. M., Millner, R., Fabri, B. M. and Jones, M. on behalf of the North West Quality Improvement Programme in Cardiac Interventions (2003) Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. *Br. Med. J.*, **327**, 13–17.

Bristol Royal Infirmary Inquiry Panel (1999) The Inquiry into the management of care of children receiving complex heart surgery at the Bristol Royal Infirmary. *Report*. (Available from `http://www.bristol-inquiry.org.uk/`.)

Bristol Royal Infirmary Inquiry Panel (2001) Monitoring standards and performance. In *Learning from Bristol: the Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary 1984-1995*. London: Stationery Office. (Available from `http://www.bristol-inquiry.org.uk/final_report/report/sec2chap30_32.htm`.)

Carter, D. (2003) The surgeon as a risk factor: determinants of outcome include technical skill, volume of work, and case-mix. *Br. Med. J.*, **326**, 832–834.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*. Belmont: Wadsworth.

Consumers' Association (2003) *Memorandum PST 46 to the House of Commons Public Administration Select Committee*. Consumers' Association, London.

Deming, W. E. (1986) 14 points for management. In *Out of the Crisis*. Cambridge: Massachusetts Institute of Technology Press.

Department of Health (2002) Methodology—acute NHS hospital trusts. Department of Health, London. (Available from http://www.doh.gov.uk/performanceratings/2002/method_acute.html.)
Department of Health (2003) Methodology—performance ratings. Department of Health, London. (Available from http://www.doh.gov.uk/performanceratings/2003/acute_details.html).
Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (2002) Is more information better?: the effects of 'report cards' on health care providers. *Working Paper w8697*. National Bureau of Economic Research, Cambridge.
Dyer, O. (2003) Heart surgeons are to be rated according to bypass surgery success. *Br. Med. J.*, **326**, 1053.
European Union (1997) Council Regulation 322/97 on community statistics, Feb. 17th. European Union, Brussels. (Available from http://www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/legislation/stat_law.html.)
Fairweather, C. (2002) Independent inspection—safeguarding the public, promoting correctional excellence. In *Prisoners under Scrutiny: a Human Rights Perspective*. Dublin: Irish Penal Reform Trust. (Available from http://www.penal-reform.ie.)
Fuggle, S. V., Belger, M. A., Johnson, R. J., Ray, T. C. and Morris, P. J. (1999) A new allocation scheme for adult kidneys in the United Kingdom. In *Clinical Transplants 1998* (eds J. M. Cecka and P. I. Terasaki), pp. 107–113. Tissue Typing Laboratory, University of California at Los Angeles, Los Angeles.
Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.
Gore, S. M. and Drugs Survey Investigators' Consortium (1999) More effective monitoring needed of young people's use of illegal drugs: meta-analysis of UK trends. *Br. J. Criminol.*, **39**, 575–584.
Gray, J., Goldstein, H. and Thomas, S. (2001) Predicting the future: the role of past performance in determining trends in institutional effectiveness at A level. *Br. Educ. Res. J.*, **27**, 391–406.
Hawkes, N. (2003) Dismissed NHS chief awarded £280,000. *The Times*, June 5th, 5.
Her Majesty's Government (1998) *Statistics: a Matter of Trust*. London: Stationery Office.
Her Majesty's Government (1999) *Building Trust in Statistics*. London: Stationery Office.
Her Majesty's Government (2003) Health and Social Care (Community Health and Standards) Bill—draft. London, Her Majesty's Government. (Available from http://www.publications.parliament.uk/pa/ld200203/ldbills/094/2003094.htm.)
Her Majesty's Treasury, Cabinet Office, National Audit Office, Audit Commission and Office for National Statistics. *Choosing the Right Framework: a Framework for Performance Information*. London: Stationery Office. (Available from http://www.hm-treasury.gov.uk/documents/public_spending_and_.)
Higher Education Funding Council for England (2002) *Performance Indicators in Higher Education in the UK, 1999-2000, 2000-2001*. Bristol: Higher Education Funding Council for England.
House of Commons Public Administration Select Committee (2002a) Examination of Witnesses (Professor Tim Brighouse, Mr David Butler, Director, National Confederation of Parent Teacher Associations, Dr Gill Morgan, Chief Executive, NHS Confederation and Mr Mike Newell, President, Prison Governors' Association), Dec. 5th. *Questions 362–451*.
House of Commons Public Administration Select Committee (2002b) Examination of Witnesses (Mr Roger Thayne OBE, Chief Executive, Staffordshire Ambulance Service, Councillor Sir Jeremy Beecham, Chairman, and Mr Matthew Warburton, Head of Futures, Local Government Association, and Mr Mike Stone, Chief Executive, Patients Association), Nov. 21st. *Questions 142–230*.
House of Commons Public Administration Select Committee (2003a) Examination of Witnesses (Professor Michael Barber of the Prime Minister's Delivery Unit and Mr Nick Macpherson of Public Services Directorate in Her Majesty's Treasury), Feb. 27th. *Questions 486–614*.
House of Commons Public Administration Select Committee (2003b) *On Target?: Government by Measurement*, vol. 1. London: Stationery Office. (Available from http://www.publications.Parliament.uk/pa/cm/cmpubadm.htm#reports.)
House of Commons Public Administration Select Committee (2003c) Examination of Witnesses (Mr James Strachan and Mr Peter Wilkinson from Audit Commission), Jan. 9th. *Questions 486–614*.
House of Commons Public Administration Select Committee (2003d) Examination of Witnesses (Mr Peter Neyroud, Chief Constable, Thames Valley Police, Mr Martin Narey, Director General, Prison Service, Professor Alison Kitson, Executive Director (Nursing), Royal College of Nursing (RCN) and Mr John Seddon, Managing Director, Vanguard Education Limited), Jan. 23rd. *Questions 677–765*.
Jackson, G. W. L. (2002) Drug-related deaths in Scotland in 2001. General Register Office for Scotland, Edinburgh. (Available from http://www.gro-scotland.gov.uk/grosweb/grosweb.nsf/pages.)
Jacobson, B., Mindell, J. and McKee, M. (2003) Hospital mortality league tables: question what they tell you—and how useful they are. *Br. Med. J.*, **326**, 777–778.
Klein, S. P., Hamilton, L. S., McCaffrey, D. F. and Stecher, B. M. (2000) What do test scores in Texas tell us? *Educ. Poly Anal. Arch.*, **8**, 1–21.
Moore, P. G. (Chair) (1991) Official statistics: counting with confidence. *J. R. Statist. Soc.* A, **154**, 23–44.
Office for National Statistics (2000) *Framework for National Statistics*. London: Office for National Statistics.
Office for National Statistics (2002) *Code of Practice*. London: Office for National Statistics.

Propper, C. and Wilson, D. (2003) The use and usefulness of performance measures in the public sector. *Oxf. Rev. Econ. Poly*, **19**, 250–267.
Roberts, G. (2003) Review of research assessment: report to UK joint funding bodies. *Report*. Wolfson College, Oxford.
Smith, A. F. M. (1996) Mad cows and ecstasy: chance and choice in an evidence-based society (the Address of the President). *J. R. Statist. Soc.* A, **159**, 367–383.
Smith, J. (Chair) (2001) The Shipman Inquiry. (Available from `http://www.the-shipman-inquiry.org.uk/home.asp`.)
Smith, P. (1995) On the unintended consequences of publishing performance data in the public sector. *Int. J. Publ. Admin.*, **18**, 277–310.
Spiegelhalter, D. J. (1999) Surgical audit: statistical lessons from Nightingale to Codman. *J. R. Statist. Soc.* A, **162**, 45–58.
Spiegelhalter, D. (2002) Funnel plots for institutional comparison. *Qual. Safty Hlth Care*, **11**, 390–391.
Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. W. and Murray, G. D. (2002) Commissioned analysis of surgical performance using routine data: lessons from the Bristol Inquiry (with discussion). *J. R. Statist. Soc.* A, **165**, 191–232.
Statistics Commission (2003a) *Memorandum PST 21 to the House of Commons Public Administration Select Committee*. Statistics Commission, London.
Statistics Commission (2003b) *Annual Report 2002-2003*. London: Stationery Office. (Available from `http://www.statscom.org.uk/resources/reports.htm`.)
Steiner, S. H., Cook, R. J. and Farewell, V. T. (1999) Monitoring paired binary surgical outcomes using cumulative sum charts. *Statist. Med.*, **18**, 69–86.
United Nations (1994) *UN Fundamental Principles on Official Statistics*. New York: United Nations.
World Health Organization (2002) *Report of the Scientific Peer Review Group on Health Systems Performance Assessment*. Geneva: World Health Organization.
Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999) The use of assessment data for school improvement purposes. *Oxf. Rev. Educ.*, **25**, 469–483.