Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: http://www.tandfonline.com/loi/tjor20

The signal in the noise: Robust detection of performance "outliers" in health services

Nathan C. Proudlove, Mhorag Goff, Kieran Walshe & Ruth Boaden

To cite this article: Nathan C. Proudlove, Mhorag Goff, Kieran Walshe & Ruth Boaden (2018): The signal in the noise: Robust detection of performance "outliers" in health services, Journal of the Operational Research Society, DOI: 10.1080/01605682.2018.1487816

To link to this article: https://doi.org/10.1080/01605682.2018.1487816

Published online: 18 Oct 2018.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
https://doi.org/10.1080/01605682.2018.1487816

The signal in the noise: Robust detection of performance "outliers" in health services

Nathan C. Proudlove, Mhorag Goff, Kieran Walshe and Ruth Boaden

Alliance Manchester Business School, University of Manchester, Manchester, UK

ABSTRACT

To make the increasing amounts of data about the performance of public sector organisations digestible by decision makers, composite indicators are commonly constructed, from which a natural step is rankings and league tables. However, how much credence should be given to the results of such approaches? Studying English NHS maternity services ($N = 130$ hospital trusts), we assembled and used a set of 38 indicators grouped into four baskets of aspects of service delivery. In the absence of opinion on how the indicators should be aggregated, we focus on the uncertainty this brings to the composite results. We use a large two-stage Monte Carlo simulation to generate possible aggregation weights and examine the discrimination in the composite results. We find that positive and negative "outliers" can be identified robustly, of particular value to decision makers for investigation for learning or intervention, however results in between should be treated with great caution.

ARTICLE HISTORY
Received 31 October 2016
Accepted 1 June 2018

KEYWORDS
Health service; hospitals; management; simulation; quality

1. Introduction

Healthcare systems, and their failures, have a very high profile in public and political consciousness. There are many groups with an interest in assessing the performance of organisations providing healthcare: the public as consumers and (directly or indirectly) as funders; patient advocate groups, charities and the media; health professionals and the provider organisations' own staff; and, most directly, those with a surveillance, regulatory and/or policy-making role (whom we will term here "decision makers"). These groups present a challenge to analysts, continually calling for more data to be made available whilst desiring an easily-digestible overview, avoiding the need to delve into the detail and context, but without clear and consistent views of how this should be done (Pidd, 2012). A particular goal of decision makers is to find signals of outliers: best-performing organisations (for reward or "best-practice" learning) and the worst (for punishment or intervention/assistance).

Investigations into a succession of major patient-safety scandals in the UK NHS over the last three decades call for more and better data and communication. However, the volume of data is not necessarily the issue. Several of these public inquiries use terms like "awash with data" (Kennedy, 2001, p. 3; Macrae, 2014), criticising organisations for their inability to make use of data, but nevertheless recommending the collection of yet more. Burgess (2012) notes the very large number of "quality" performance indicators in use in the NHS, which has been characterised as a "hypercomplex" system (Klein & Youngng, 2015) with multiple goals and stakeholders. Macrae (2014) points out that having so much data makes it harder to spot signals, especially "weak signals" (Ansoff, 1975), of major problems in healthcare systems. The NHS regulator, the Care Quality Commission, is itself wrestling with how to identify potential risks from these big datasets (Bardsley et al., 2009), as yet unsuccessfully (Griffiths, Beaussier, Demeritt, & Rothstein, 2017). Such issues are not confined to the UK. For example, in Australia an investigation into excess perinatal deaths (Wallace, 2015) led to the dismissal of a health services board and criticism over failure to detect problems much earlier (Davey, 2015).

In this environment the use of composite indicators, rankings and league tables for easy consumption by decision makers is very common and appears inevitable (Jacobs & Goddard, 2007; Talbot, 2010). This is despite criticisms from academics about issues such as how aggregation weightings are chosen, the effects of uncertainty in the underlying data and, in particular, evidence that resulting

CONTACT Nathan Charles Proudlove nathan.proudlove@manchester.ac.uk Alliance Manchester Business School, University of Manchester, Manchester M13 9SS, UK.

league tables tend to have little discrimination in mid-ranges, see Goldstein and Spiegelhalter (1996), Jacobs, Goddard, and Smith (2005) and Leckie and Goldstein (2009). As Pidd (2012) concludes, apparent differences in performance must be approached with great care, particularly in the mid-reaches of league tables where the effects of variation tend to be greatest, though they may be of use in identifying the extremes, that is their real value may be to provide actionable signals of outlying performance. This is apparent in the confidence-interval approach to aggregate scores produced during examining the robustness of ranking of schools (Leckie & Goldstein, 2009) and hospitals (Jacobs & Goddard, 2007), though this is not a point these authors focus on.

Research into rankings from composite indicators has generally focused on uncertainty about values of the data, making parametric assumptions about the nature of the data-generation mechanisms and searching for statistically-significant outliers, whether as the signal to be identified or as extreme noise to be corrected for (e.g., Spiegelhalter et al., 2012). Our approach in this paper is to avoid such assumptions which continue to be critiqued (Mingers, 2006; Ziliak & McCloskey, 2008) and which we argue would be artificial with the sort of situation we are interested in. Instead we focus on the effects of different weightings that could be used in aggregating the indicators to form a composite score. Therefore, rather than statistical significance, we focus on robustness and size of effects.

The work reported in this paper focuses on the nationally-available performance indicators for maternity services in NHS acute hospital trusts. Maternity is a particularly high-profile domain, and one in which a large number of indicators has recently become available, spanning clinical processes, mortality, regulator judgements and patient experience, but with no consensus on how these might be aggregated. The key effect of uncertainty that we are concerned with here is the impact of possible sets of weights.

Our research questions are:

- RQ1. Do the indicators suggest that trusts can be ranked meaningfully?
- RQ2. Do the indicators suggest that there are clear high- and low-performing trusts?

We assembled data from a wide range of sources, the first time such a comprehensive set has been assembled for maternity; previous studies have focused on particular aspects such a clinical indicators (Carroll, Knight, Cromwell, Gurol-Urganci, & van der Meulen, 2016) or mortality (Manktelow et al., 2015). We first use population (rather than sample) focused statistical techniques to look for structure across the indicators, but the core technique is a two-level Monte Carlo simulation to perform a probabilistic sensitivity analysis of the effects of different indicator weightings. Thus we are taking the opportunity for operational research (OR) to contribute to this challenge of searching for robust meaning in complex datasets with approaches from the simulation and data mining facets of the discipline (Baesens, Mues, Martens, & Vanthienen, 2009). Indeed Royston (2013) lists "accommodating analytics" as one of the challenges for "OR". OR can bring perspectives and techniques beyond econometric-type multivariate statistics which, as Mingers (2006) points out, can be applied with naïve epistemology and very poor predictive performance. Previously, simulation has been used to examine the effect on rankings of random variation (noise) in the individual indicators (Jacobs & Goddard, 2007; Marshall & Spiegelhalter, 1998) but the suggestion to apply it to the aggregation weights (Spiegelhalter et al., 2012, Expert discussion on the paper) does not appear to have been taken up.

Our results clearly and transparently exhibit an important case of where ranking-position is not very meaningful for "middling" organisations but the extremes can be identified robustly. The message for decision makers is that ranking organisations through composite indicators is dangerous, but potentially valuable for identifying these outliers. This is an interesting situation in which the outliers are the signal in data rather than being distorting noise. We demonstrate the use of simulation, a different and more sophisticated method than applied previously, to combine multiple sources of data and robustly handle uncertainty about the relative weightings of a large number of performance indicators.

2. Outliers in data

The term "outlier" is used differently in different disciplines and contexts, and consequently appropriate reactions to their presence in data differ. In mainstream statistics an outlier in data often represents a problem. It is generally seen as a rogue, unrepresentative observation, contaminating a dataset and "reducing and distorting the information about the data source or generating mechanism" (Barnett & Lewis, 1994, p. 4); it "distorts statistics" (Tabachnick & Fidell, 2013, p. 72) about the underlying "regular" population of interest. Similarly, in the modelling fields spanned by OR, outliers tend to be regarded as disruptive noise in empirical data, inhibiting our estimation of parameters for models in, for example: regression modelling (e.g., Hindle, Hindle, & Souli, 2009); demand analysis (e.g.,

Banerjee & Meitei, 2010; Zhao, Zhao, & Shen, 2017) forecasting (e.g., Trapero, Kourentzes, & Fildes, 2015) and model selection (e.g., Tofallis, 2015); simulation (e.g., Hoad, Robinson, & Davies, 2010); credit risk scoring (e.g., Florez-Lopez, 2010); efficiency frontier analysis (e.g., Daraio & Simar, 2016); and case classification (e.g., Yan, Bai, Fang, & Luo, 2016). Common methodologies are to seek to detect outliers, then consider remedies such as segmenting the data, accommodating them through "robust" methods, or deleting, replacing or down-weighting such data points (Ord & Fildes, 2013; Tabachnick & Fidell, 2013).

In contrast, there are situations where "outliers" represent some or the important information: valuable signals of unexpected or different (usually interesting) behaviour from the target system under investigation. At the research strategy level, outliers ("anomalous" or "deviant" cases) are crucial in the deductive part of cycles of theory testing and building according to the scientific method, refutation of hypotheses signalling the need for new and better theory (Christensen, 2006; Flyvbjerg, 2011). In data-driven methods (for example machine learning or data mining), the detection of outliers (commonly termed "anomalies") can also be the goal, detecting cases of interest for example potential financial fraud or equipment fault (Chandola, Banerjee, & Kumar, 2009). Howell and Proudlove (2007) found major outliers in fits of multiple-regression models to very noisy retail data picked out cases of fraud or gross mismanagement by store managers. In healthcare, Harley, Mohammed, Hussain, Yates, and Almasri (2005) and Mohammed, Cheng, Rouse, and Marshall (2001; 2004) show how statistical detection of outliers in routinely-collected NHS performance data might have alerted authorities to medical malpractice many years earlier, including the case of the mass-murdering general practitioner Harold Shipman. In OR, an outlier may be a divergent individual in a group decision support process who may have a valuable perspective or resist dangerous groupthink in problem structuring (Shaw, Westcombe, Hodgkin, & Montibeller, 2004) or "special cause variation" that needs to be investigated to learn how further occurrences might be designed in or out of a process (Pidd, 2012; Wood, Capon, & Kaye, 1998).

There are conventional (or "frequentist") statistical approaches to the detection of outliers, for example the Mahalanobis distance (Gnanadesikan, 1997) and funnel-plots are now often recommended to pick out potentially meaningful outliers (Bird, Farewell, Goldstein, Holt, & Smith, 2005). These rely on assumptions about the distribution of parameters, sometimes after scale transformation, under null hypotheses in order to calculate p-values. However, assumptions about distributions may be without empirical justification, transformations may be perceived as opaque and unfair, and the distributions of data can mask outliers from p-value detection based on the Mahalanobis distance (Spiegelhalter et al., 2012, Expert discussion on the paper).

3. Methodology

The work described in this paper is part of a wider project funded by the Health Foundation in the UK. The project seeks to understand the relationships between structures, processes and capabilities in NHS healthcare providers and their ability to deliver clinical quality and performance improvement (Darley, Walshe, Boaden, Proudlove, & Goff, 2018). This wider project uses a multimethod approach. The quantitative modelling components described in this paper supported the early stages of the project, examining patterns and structure in the publically-available data. The research questions addressed in this paper informed where to look (at what performance aspects and which trusts) for the qualitative fieldwork case investigation of underlying causal mechanisms.

The collection and reporting of indicators are subject to complex mechanisms which add "noise" to any underlying signal of actual performance. Attempts have been made to estimate the extent of this noise in these types of data by assuming it is random variation and using regressions on longitudinal (3 years plus) data series (resulting in estimates from 1% to 98%), then assuming all the organisations are subject to this randomness drawn from a common normal probability distribution (Jacobs et al., 2005; Jacobs & Goddard, 2007).

In our case, few of our indicators are regularly and systematically captured, processed and published, so we cannot attempt to do this. However there are also other reasons we chose not to go down this type of statistical route. While there may well be mechanisms involved in the generation of some indicators in some organisations which in aggregation do look like simple natural randomness, other indicators are subject to policy control with different choices being made in different organisations. So assuming a common mechanism seems hard to justify, as would be assuming smooth, unimodal error probability distributions (normal, Poisson, etc.) to enable the calculation of p-values and confidence intervals.

A deeper objection is that the above approach assumes that a set of indicators (e.g., for a year) is a sample (but with $n = 1$) drawn from a "population" of possible values with inferred characteristics. This

Table 1. Data sources.

Indicator set	Source	Summary/context	Period
NHS Maternity Statistics	NHS digital (ex-Health and Social Care Information Centre (HSCIC)) http://digital.nhs.uk/catalogue/PUB19127/nhs-mate-eng-2014-15-tab.xlsx	• Derived from Hospital Episode Statistics (HES) data • Characteristics of the birth including method of onset, delivery method, types/rates of intervention and unassisted delivery.	2004 onwards annual Used: 2014–2015
RCOG Clinical Indicators Project	Royal College of Obstetricians and Gynaecologists (RCOG) https://indicators.rcog.org.uk/ (Carroll et al., 2016)	• Derived from HES data • Suite of risk-adjusted indicators for intrapartum care processes and selected adverse outcomes, intended to define a standard for NHS maternity services in England.	annual from 2015 Used: 2016 publication (derived from 2013–2014 data)
MBRRACE-UK	Mothers and Babies: Reducing Risk through Audits and Confidential Enquiries across the UK collaboration (MBRRACE-UK) National Perinatal Epidemiology Unit, University of Oxford https://www.npeu.ox.ac.uk/mbrrace-uk (Manktelow et al., 2015)	• Mandatory bespoke data collection and reporting by trusts • Report based on 2013 data • Stillbirth, neonatal and extended perinatal mortality rates	2015 Used: 2015 publication (derived from 2013 data)
CQC Inspection Reports	Care Quality Commission (CQC) www.cqc.org.uk/content/how-get-and-re-use-cqc-information-and-data#directory	• Inspections of provider organisation sites in key service areas • Indicators rate performance from Outstanding to Inadequate across five dimensions (Safe, Effective, Caring, Responsive, Well-led) plus Overall	Rolling programme of inspections and reporting 2009–2016
CQC Survey of Maternity Service Users	CQC www.cqc.org.uk/content/maternity-services-survey-2015	• Survey of women who used maternity services in England over 1 or 2 months.	2010, 2013 2015 Used: 2015

focus on sampling error, with necessary distributional assumptions or transformations, then typically takes statisticians down the route of statistical significance as the prime criterion for detecting unusual performance (e.g., Spiegelhalter et al., 2012). The misuse in many fields of such chains of logic leading to "sizeless" conclusions (ignoring the importance of material significance), and even wrong or non-robust (e.g., non-replicable) "knowledge," is demonstrated and critiqued thoroughly by Ziliak and McCloskey (2008). Mingers (2006) also draws together the many issues in the (mis-) use of inferential statistics, with particular reference to applications in OR.

Since we have datasets covering *all* providers (with the caveats below), we regard our data as a population, rather than a sample with some assumed error characteristics. We therefore are using descriptive statistics and inductive structure-finding techniques, rather than conventional inferential statistics based on parametric models from sampling theory. The motivation behind the work is to look for large and robust variation in performance between the hospital trusts which may indicate different causal behaviours and so inform subsequent qualitative fieldwork investigation (the sort of multimethod approach advocated by Mingers, 2006). There is therefore an implicit assumption that patterns found from these (recent) historical data say something about current performance. A recent study of NHS trusts' performance over many years in a different activity area (short-notice cancelled elective operations) demonstrates both that variation between trusts is large and that relative performance is highly persistent over time (Proudlove, Samarasinghe, & Walshe, 2018).

In summary, a key feature of our approach is the avoidance of the statistical assumptions and approaches commonly used in performance assessment. Our approach is, though, not completely assumption-free, and we will use the shorthand "assumption-light".

3.1. Maternity performance indicator data

We chose to look at maternity services since (i) they are relatively self-contained, isolated in many ways from wider issues affecting the performance of a hospital trust, (ii) they are important, accounting for a disproportionately large and rising share of the NHS's litigation costs (National Audit Office, 2013), (iii) there is a relatively large number of performance measurement indicators becoming available, and (iv) they are particularly high-profile, including the recent Morecambe Bay maternity care scandal (Kirkup, 2015) and concern over unexplained performance variation (e.g., Carroll et al., 2016).

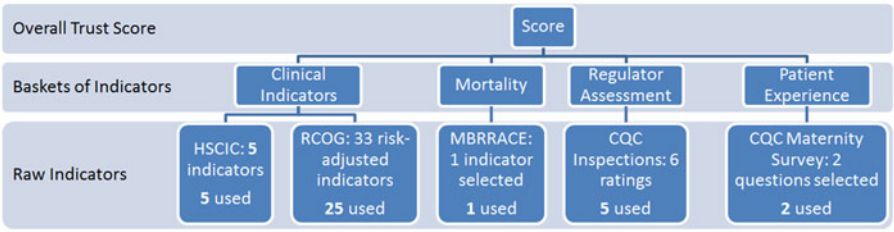

Figure 1. Hierarchy of indicators.

Maternity data are collected and published for a range of purposes, and much has become available for the first time very recently. Table 1 lists the sources from which we assembled our dataset.

There are many more indicators on some aspects than others. To prevent some aspects drowning out others we aggregate them hierarchically into four meaningful "baskets" (Spiegelhalter et al., 2012) as shown in Figure 1.

Despite the effort put into collecting, cleaning, correcting and filtering these national datasets, all sources acknowledge the variable and uncertain quality of the data (e.g., Carroll et al., 2016). Our approach here is to not regard "outliers" in our population as rare chance outcomes from distributions of (specifiable) randomness or invalid data points, but as real members of the population, though taking a robust approach, as detailed below.

As common, (e.g., Harley et al., 2005; Proudlove, 2012; Spiegelhalter et al., 2012) we use Z-scoring, so that each indicator has a mean of zero and a standard deviation of one, to make them comparable, and with a high Z-score being "bad" (an indicator worse than the mean); where appropriate, indicators are reverse scored. We note that the clinical indicators are not all straightforwardly directional (in the sense that a higher or lower value can be definitively regarded as better or worse). For example, assisted delivery is of course sometimes required for safety, but is also associated with risk of harm so should only be used when necessary; consequently there has been concern about increased rates of intervention without corresponding improvements in maternal or neonatal outcomes, suggesting over-intervention (The Royal College of Midwives, 2007; The King's Fund, 2008). Similarly, there is concern about the rise in the caesarean section rate in the UK without evidence of net benefit in terms of outcomes and without evidence of the impact on the long-term consequences for maternal health (Savage, 2000). CQC inspection reports sometimes praise initiatives to reduce rates of intervention in particular maternity departments (e.g., Care Quality Commission, 2015). We argue, therefore that, in the range of values, higher Z-values generally indicate worse practice.

We noted the frequent caveats from the sources about data quality issues; further, many of the indicators are derived from ratios (rates of occurrences) which can inflate the spread in data. We therefore take the simple, conservative and robust approach, as is common in such circumstances (Baesens et al., 2009; Spiegelhalter et al., 2012), of Winsorizing the Z-scored indicators (constraining them to ±3) and Z-scoring again to restore centring at zero and spread of one. In practice we found this made little difference to the results.

Additional contextual factors about trusts are maternity volume, MBRRACE Table (risk) category and Unit level (intensity of service), and a measure of complex social factors (CSFs; ethnic and deprivation factors known to influence maternity outcomes; Carroll et al., 2016) but only available for about a third of the cases.

3.2. Aggregating indicators

Techniques to deal with multiple indicators, which may be aggregated hierarchically, should be straightforward, easy to explain to a range of stakeholders and robust to potential issues with data quality (Spiegelhalter et al., 2012). Linear additive models, with weights scaled to sum to one, are straightforward and transparent. Several approaches have been suggested to setting the weights: normative weights (based on expert judgement or surveys) perhaps calculated from multi-attribute decision making techniques such as pairwise comparison; equal weighting; or, since correlations between indicators can result in more weight being put on one aspect than intended (or apparent), setting weights accounting for this, for example derived from conjoint analysis, principal components analysis (PCA) or down-weighting positively correlated variables (Pidd, 2012; Spiegelhalter et al., 2012). Other simple aggregation systems include decision rules, as used by the CQC in aggregating its five domain ratings to an Overall rating, which can have more face validity for decision makers or clinicians but the results are particularly sensitive to the rules and thresholds imposed (Jacobs et al., 2005). More complex, so less transparent to stakeholders, methods include "benefit-of-the-doubt": setting each unit's weights to give it the highest relative score, subject to constraints; a special case of data envelopment analysis (Karagiannis, 2017).

Table 2. Correlations between the four baskets (bold) and contextual variables.

		Clinical Indicators	Mortality	Regulator assessment	Patient experience	Total score	Number of maternities	Highest unit level	MBRRACE table	Complex social factors
Baskets	Clinical indicators		**−0.305**	**0.134**	**−0.068**	0.409	0.112	0.087	−0.063	0.161
	Mortality	**−0.305**		**−0.202**	**0.013**	0.272	−0.061	0.016	−0.011	0.291
	Regulator Assessment	**0.134**	**−0.202**		**0.157**	0.586	0.016	−0.093	0.076	−0.003
	Patient experience	**−0.068**	**0.013**	**0.157**		0.593	0.238	0.099	−0.034	0.146
	Total score	0.409	0.272	0.586	0.593		0.164	0.059	−0.017	0.329
Contextual	Number of maternities	0.112	−0.061	0.016	0.238	0.164		0.611	−0.758	0.177
	Highest unit level[a]	0.087	0.016	−0.093	0.099	0.059	0.611		−0.844	0.280
	MBRRACE risk table[b]	−0.063	−0.011	0.076	−0.034	−0.017	−0.758	−0.844		−0.325
	Complex social Factors	0.161	0.291	−0.003	0.146	0.329	0.177	0.280	−0.325	

$N = 75$ trusts.
[a]1: Local Neonatal Unit (LNU); 2: Special Care Baby Unit/Special Care Unit (SCU); 3: Neonatal Intensive Care Unit (NICU).
[b]see Table 1. MBRRACE splits trusts into five tables according to broad risk categories (based on neonatal surgical provision, NICU and number of maternities).

However indicators are aggregated though with weighting or rule structures there is often little consensus about how these should be determined, and the results are sensitive to the methodological choices made (Jacobs & Goddard, 2007). For maternity, there is no expert or normative model for the weightings, as confirmed by our project's small reference group of expert practitioners and clinical academics. Midwives and obstetricians tend to regard birth through different lenses, which give them different professional framings of quality (Graham & Oakley, 1986; MacKenzie Bryers & van Teijlingen, 2010). There has been a lack of consistency in the quality indicators being used (Boulkedid, Alberti, & Sibony, 2013), which explicitly motivated RCOG's own clinical indicators project (Carroll et al., 2016), which deliberately produced a wide set, which we use in this study.

Spiegelhalter et al. (2012, Expert discussion on the paper) contains a suggestion to use Monte Carlo simulation to examine the effects of different weights, for example in the range between equal weights and expert (e.g., clinical) judgement. Here, without a steer from such expert judgements, we allow the simulation to pick weights in the range (0,1) for each indicator (which are then scaled to sum to one). In previous work healthcare performance indicator analysis, simulation has been used to examine the possible effects of uncertainty (measurement and random error), with weights only varied deterministically to construct a few scenarios (Jacobs & Goddard, 2007). Brailsford, Harper, Patel, and Pitt (2009)'s review of healthcare modelling found Monte Carlo simulation to be a fairly common approach, but overwhelmingly used as a secondary technique to a primary method, often for probabilistic sensitivity analysis. Here this is our primary technique.

The basket structure (Figure 1) was constructed to allow equal prominence to the four main aspects of performance covered, so we use a two-level simulation design. We use Monte Carlo sampling to generate weights for the indicators within each basket, scaling the weights to sum to one. This is repeated at the higher level, weighting the four baskets to give an overall possible score for each trust. So, for each of n simulation trials a vector of results is produced for the set of t trusts, so this produces a $t \times n$ matrix for each basket. The four basket matrices are then stacked in a third dimension, to form a $t \times n \times 4$ array (a tensor of rank 3). The higher level of the simulation generates weights for each basket on each trial in the form of a $4 \times n$ matrix. This is then multiplied along the trials dimension to produce a $t \times n$ results matrix from which the final results are derived. We used the *R* high-level programming language (The R Foundation), version 3.3.1 for our analyses. In *R* these operations can be coded very succinctly using array and tensor functions and executed very quickly for example <20 s for 100,000 trials on a not particularly powerful PC. With this number of trials, different random number generator seeds produced no material difference in the final results. *R* is also very powerful for designing customised graphical output.

4. Findings

4.1. Initial analysis

After excluding trusts that have been dissolved or undergone very major reorganisations over the timeframe of our dataset, we had data for 130 trusts. 55 had some missing data, all this being in the clinical indicators basket. Some of these resulted from the HSCIC's judgements that figures from HES data (submitted by trusts themselves) were too limited or of insufficient quality for the HSCIC to publish. Overall, 8% of the indicator values were missing. With generally low correlations between indicators, and some trusts missing quite a large number of data points, we did not attempt imputation. For the simulation we ran both the complete data only ($N = 75$) and all trusts ($N = 130$) with missing values replaced with *Z*-scores of zero: a neutral contribution.

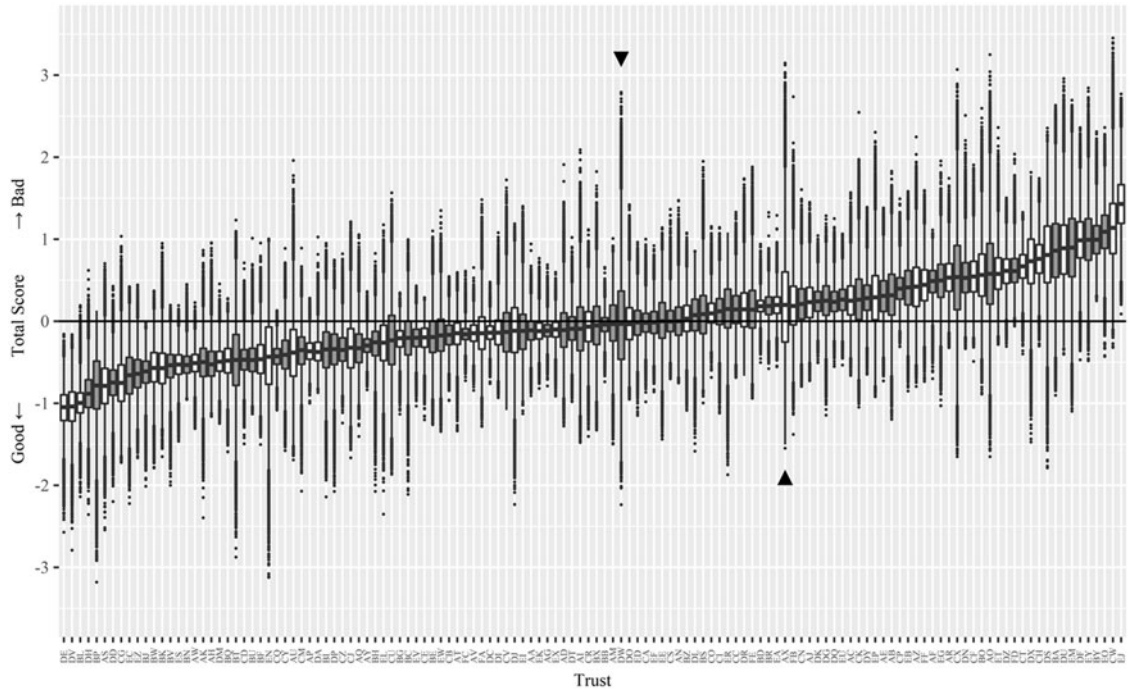

Figure 2. Boxplots of trust total scores from Monte Carlo Simulation; 100,000 trials; ordered by mean score. $N = 130$ trusts. Unshaded boxes indicate trusts with incomplete data.

Processes and outcomes in healthcare are obviously affected by the clinical severity of patient cases (case-mix), and larger (higher-volume) services also tend to have facilities to treat higher-intensity cases. The RCOG data have some degree of risk-adjustment, though they note some maternal risk factors were not available to them (Carroll et al., 2016). Where the same indicator is present in the RCOG and HSCIC sources we found little difference in value. The only major issue with risk-adjustment found in our dataset was Mortality. MBRRACE presented mortality data stratified into five tables (groups) based on the intensity of service provision and volume of cases; the differences in mean mortality rates between tables are material. Therefore, for this study we Z-scored the mortality data separately within each table, with the two lowest-risk tables combined since the smallest contained few trusts.

Considering just the complete cases ($N = 75$ trusts), an initial correlation sweep of the 47 indicators revealed some very strong intercorrelations ($r > 0.8$): some variables being essentially proxies for others. To remove this collinearity, nine indicators (eight RCOG and one CQC Inspection) were removed, resulting in the set of 38 across four baskets used (Figure 1). Excluding trusts with incomplete data where necessary, we used PCA to look for structure within the indicators both across all variables and within each basket. The structures are weak, with little useful opportunity to collapse the 38 variables into fewer components, which would anyway come at the expense of much-reduced interpretability.

To examine correlations between the baskets, average scores were calculated by equally weighting the sets of metrics within each. Table 2 shows the correlations between these average basket scores plus with an equally-weighted (averaged) total score and also with the contextual variables. The Mortality ↔ Clinical Indicators correlation just reaches "medium" strength ($|r| > 0.3$). This may be a result of residual case-mix risk. As noted above, there is risk adjustment in some of the indicators for the type of maternity provision in a trust, but only for some aspects of factors to do with the mother and baby. The other correlations between the baskets and with contextual variables are weak, so overall we have four essentially independent dimensions. The lack of correlation between the Regulator Assessment and the other three baskets is interesting. It suggests that, though the CQC Inspection teams have access to performance indicator data, they may not make much use of it, or that the qualitative data and first-hand observations gathered during the inspection visits may dominate their assessments. Similar conclusions were reached by studies of the inspection process (Walshe, Addicott, Boyd, Robertson, & Ross, 2014) and trying to relate outcomes to sets of indicator data considered particularly relevant in particular service areas by the CQC (Griffiths et al., 2017).

4.2. Simulation results

Figure 2 shows the results of using simulation to generate many different sets of weightings of the 38

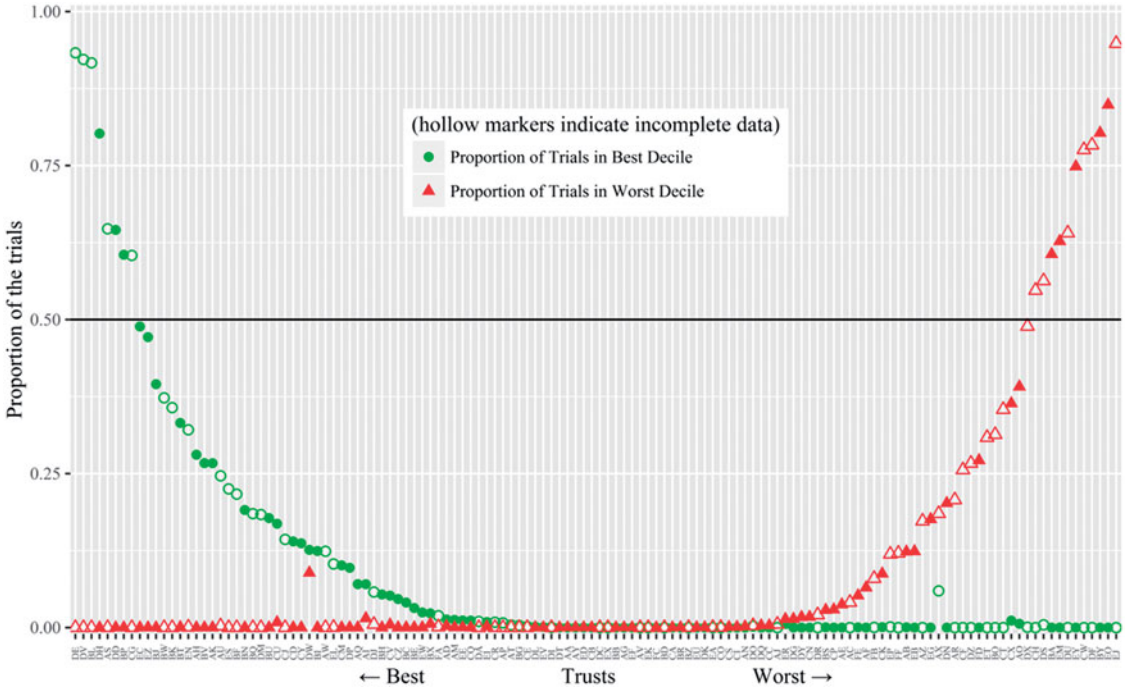

Figure 3. Proportion of trials (sets of weights) in which a trust has a total score in the top and bottom decile from Monte Carlo simulation, 100,000 trials; trusts ordered by highest proportion, $N = 130$ trusts.

indicators and four baskets. Essentially this is examining how sensitive the trusts' overall scores are to different weighting possibilities.

Figure 2 shows the great variability in total score possible. Only two trusts come out better than average (total score <0) and one worse (total score >0) in every trial. Trusts with widely-varying scores across individual indicators have a wide range of total score. In particular the trust indicated by the downward pointing triangle has a "tail" of high (bad) total scores, sufficient to frequently produce a higher (worse) score than many of its neighbours when simply arranged by mean (as on Figure 2). In fact it has good performance on Clinical Indicators, Regulator Assessment and Patient Experience, but *very* poor performance on Mortality. When mortality is relatively-highly weighted in a trial, the trust comes out higher (worse) than its neighbours. The other trust picked out (upward-pointing triangle) has very poor performance on Patient Experience, but is good on the other three baskets, resulting in a tail of low (good) scores relative to its neighbours.

Given the variability in trusts' scores under different weighting systems (Figure 2), we used the percentage of trials in which a trust appeared in the top or bottom deciles (top or bottom 13 trusts) as an overall assessment of its performance. This is shown in Figure 3.

We included in the simulation the 55 trusts with some indicator data missing, by replacing the missing data with *Z*-scores of zero (neutral contribution). These trusts are displayed with hollow markers in Figure 3, and unshaded in Table 3. They are scattered through the results, and only two exceed the threshold for the number of missing values accepted in previous studies with messy healthcare data (e.g., Jacobs et al., 2005).

Table 3 shows the trusts that come out in the top or bottom decile in more than 50% of the trials. These are the trusts we tentatively identify as outliers. The patterns of performance across the baskets are fairly consistent. Only for trusts BP, DU, and CW is the result dominated by one very good or very poor basket. CW might be a particular concern for the regulator (and potentially for patients) because, as well as the terrible Regulator Assessment, it has a lot of missing data on Clinical Indicators, so the roughly average score (close to zero) on this may be missing unreported poor performance on some clinical aspects. There are no trusts in these top or bottom deciles with very good/poor performance on one basket despite very poor/good performance on the others. The closest is DS, with a fairly good Regulator Assessment but poor performance across the other baskets.

The variances and covariances in the data mean that the simulation produces a selection and ordering of these "outliers" that is more interesting than just an ordering by averages. Region D and smaller (lower volume) trusts do well, but there is insufficient evidence of a pattern. Reordering Figure 3 by number of maternities does not show a relationship by size, so the results we observe are not driven by the natural tendency of smaller-volume units to have relatively higher random variability. Whilst there is a little evidence that CSFs have an impact, the data available on this are very sparse.

Table 3. Trusts identified as likely to be in the top or bottom decile of trusts across different weighting systems. $n = 100{,}000$ trials; trusts with complete indicator data shaded.

Trust	Proportion of trials in			Contextual factors		Average of indicators					Number of missing indicators
Code	Top decile (%)	Bottom decile (%)	NHS region	Highest intensity Unit	Complex social factors	Size (by number of cases)	Clinical indicators	Mortality	Regulator assessment	Patient experience	Total score	
DE	**93**	0	C	*1*	–	*Very small*	*−0.79*	*−0.60*	*−1.96*	*−1.01*	−1.09	*1*
DV	**92**	0	D	*2*	–	*Very small*	*−2.39*	*−0.79*	*−0.66*	*−0.49*	−1.08	*1*
BL	**92**	0	B	*2*	–	*Small*	*−0.37*	*−0.98*	*−1.64*	*−1.10*	−1.02	*10*
DH	**80**	0	D	*2*	*Very low*	*Small*	*−2.02*	*−0.38*	*−0.09*	*−1.16*	−0.91	*0*
AS	**65**	0	D	*1*	–	*Very Small*	*−1.37*	*−0.27*	*−1.96*	**0.33**	−0.82	*8*
DD	**64**	0	D	*2*	–	**Large**	*−1.81*	*−0.61*	*−0.66*	*0.00*	−0.77	*0*
CG	**61**	0	D	*1*	–	*Very small*	*−0.67*	*−1.76*	**0.77**	*−1.37*	−0.75	*10*
BP	**60**	0	D	*2*	*Very low*	**Large**	*−0.47*	**0.04**	*−3.00*	**0.15**	−0.82	*0*
CH	0	**55**	B	*3*	–	*Small*	**1.24**	**1.62**	*−0.20*	**0.42**	0.77	*1*
DS	0	**56**	C	*3*	–	**Very large**	**2.02**	**2.20**	*−1.64*	**0.67**	0.81	*1*
BA	0	**61**	B	*2*	–	**Large**	*−0.67*	**2.41**	*−0.66*	**2.39**	0.87	*0*
EM	0	**63**	B	*2*	–	*Small*	*−0.96*	**2.50**	*−0.66*	**2.72**	0.90	*0*
DU	0	**64**	C	*2*	*Low*	*Very small*	**0.00**	**3.09**	**0.36**	**0.13**	0.90	*1*
EY	0	**75**	C	*2*	**Very high**	*Very small*	**0.87**	*−0.57*	**1.36**	**2.48**	1.04	*0*
CW	0	**78**	B	*2*	–	**Large**	*−0.06*	**0.48**	**3.48**	**0.81**	1.18	*10*
DF	0	**78**	C	*2*	**Very high**	*Small*	**1.87**	**1.45**	**1.17**	*−0.43*	1.02	*2*
BY	0	**80**	C	*2*	–	**Very Large**	**1.03**	**0.23**	**2.22**	**0.65**	1.03	*0*
EO	0	**85**	D	*1*	–	*Small*	*−0.19*	**1.28**	**2.22**	**1.12**	1.11	*0*
EJ	0	**95**	C	*2*	–	*Very Small*	**0.56**	**0.23**	**2.22**	**2.88**	1.47	*3*

5. Discussion

As common in the public sector, collecting and using maternity data has been a persistent issue. Over two decades ago a national maternity expert group recommended the routine collection of activity-related data for the purposes of audit, comparison, and service development and planning (Department of Health, 1993) and this is still being pursued. One of the main recommendations of the recent National Maternity Review (Cumberlege, 2016) was to develop a national set of indicators to help local maternity systems track, benchmark and improve the quality of maternity services. The Government now plans for this data system to be operational and publicly available by 2018, and has launched an improvement and innovation programme to learn from best practice, investigate problems and reduce variation in performance (Department of Health, 2016). There are no recommendations on how providers and the public use and potentially aggregate the planned data for comparison.

The lack of strong evidence from our PCA of patterns amongst our 38 indicators does not help with this. It is perhaps surprising, though fits with the observation that indicators are collected for a wide variety of different primary purposes (Bardsley, 2016; Pidd, 2012) and that quality of healthcare is a highly contested concept (Boaden & Furnival, 2016).

5.1. RQ1: Do the indicators suggest that trusts can be ranked meaningfully?

Our results show that rankings must be treated with great caution, especially in the middle-ranges of "league tables". Figure 2 is more evidence for Goldstein and Spiegelhalter (1996)'s view that rank-ordering units may lead to spurious, non-robust results, and for the recommendation from the Royal Statistical Society (Bird et al., 2005) that performance measures should always be reported with consideration of the uncertainty underlying their construction. The use of league tables by organisations to accrue status, or by governments to reward organisations (for example with greater autonomy; Talbot, 2010), should be approached with care and caveats.

5.2. RQ2: Do the indicators suggest that there are clear high- and low-performing trusts?

Given the findings on RQ1, rather than absolute rankings, we constructed a metric based on the frequency of membership of the top and bottom deciles (Figure 3) to search for trusts that were robustly appearing in these deciles under the majority of

weightings trials. The trusts identified were shown in Table 3. We consider these to be "outliers," of potential interest to investigate in order to consider learning from or (for regulators) to intervene in.

Indeed the fieldwork part of the wider project conducted in-depth qualitative investigations in a sample of trusts. Inevitably, such very intensive casework can cover only a small number of the 130 trusts, and there is (of course) no objective criterion against which to evaluate the simulation results. Nonetheless, in this small field-investigation sample the classification by the simulation of the trusts into good outlier, "middling," or poor outlier does correlate with the judgemental ratings made by the fieldwork team about the trusts' improvement capability attributes (Darley et al., 2018). This lends the simulation results at least a modicum of face validity (Balci, 1998).

Other possible validation approaches could be prediction of future high-profile scandals in maternity care, though these would obviously take some long time to manifest, and replication over time. Replication would be powerful, and should be possible once comprehensive data are released more regularly. However this is not yet possible. We used the useful data available, much of it only very recently released as irregular or one-off analysis projects or programmes.

The findings on RQ1 and 2 are in line with previous research evidence from other healthcare areas that analysis of secondary data may be used to identify potentially divergent practice in healthcare (Harley et al., 2005; Mohammed et al., 2001, 2004; Mohammed, Worthington, & Woodall, 2008; Spiegelhalter, 2005; Tennant, Mohammed, Coleman, & Martin, 2007). Pidd (2012) also discusses this in the context of educational providers. Though our overall conclusion is the same, new in this article are: the application to maternity services, analysis of so many indicators, the focus on the weightings as the source of the underlying uncertainty, and the use of Monte Carlo simulation to investigate this.

5.3. Limitations

There are persistent concerns about the value and quality of performance data in healthcare, including in maternity services, most recently highlighted in the National Maternity Review (Cumberlege, 2016). Most obviously this was manifest in our data in the form of some missing indicator values. Nonetheless, the aim in this project was to investigate what could be done with the data publicly available, and avoiding parametric assumptions about the mechanisms that generated them. Imputation of missing values was not attempted since the volume was large and we did not wish to make the analysis overly technical and opaque. Due to the way the data were made public, the datasets do not all relate to the same time period, nor is it yet possible to replicate the analyses for other time periods (other than for a very small subset of the indicators). The planned national datasets should enable comparison and monitoring over time.

Though there was quite a lot of missing data, the simulation results were robust to inclusion of the trusts with incomplete data: all the outliers for the $N = 75$ complete-data trusts were part of the outlier set for the $N = 130$ all-trusts results, and the incomplete-data trusts' results were scattered through the results without pattern (Figure 3).

As there was no prior consensus about the relative importance of some indicators over others in the literature or from our Project Advisory Group, our indicator and basket weights were sampled from a uniform distribution with limits (0,1). Where there is justified belief, or some normative model is to be imposed (e.g., Proudlove, 2012), then other restrictions could be placed on the weights generated in a simulation and/or particular probability distributions used, for example simple triangular distributions around the point judgements of domain experts. Alternatively, a suggestion in Spiegelhalter et al. (2012, Expert discussion on the paper) is to sample weights from the interval between equal weights and a set suggested by clinicians.

Our results on maternity services are, of course, tentative and should not be extrapolated to other units of the hospital trusts. Application to other areas in healthcare and the public sector would be interesting, and we would expect similar results.

6. Conclusions

This article demonstrates a generic approach to separating signal from noise in situations in which there are several performance indicators. We have used an OR technique, Monte Carlo simulation, to do this more robustly than previous parametric statistical approaches. This new approach reinforces previous findings that when trying to form an overall view from this sort of multi-attribute performance data, there is often little meaningful signal in the middle-range of scores and rankings but that it may still be possible to identify high and low "outliers". The "bathtub curve" shape of Figure 3 is a characteristic of this: showing that a particular small group of organisations appear at the extremes for the majority of weighting possibilities. This is useful since it is these extremes of performance that are what is of most concern to decision makers with a direct role in oversight, regulation and

improvement (such as the CQC and NHS Improvement). Previous work in the literature has focused on statistical uncertainty about raw indicator values, and used either fixed aggregation weights or a few sets of weights to produce scenarios. Here we have taken an assumption-light approach and focused on the effect of uncertainty about the weights used to construct a composite indicator in a situation where no normative or expert-judgement weighting model exists.

The *R* software proved a powerful tool for compact coding, handling the two- and three-dimensional data arrays involved, rapid execution of a very large number of Monte Carlo trials, post-simulation results processing, and designing customised graph formats.

Statement of contribution

The performance of public services is of great concern to governments, the public and the provider organisations themselves. This is particularly true of public health services. In the UK there has been a series of high-profile healthcare scandals, only detected after considerable avoidable death and suffering, resulting in a string of major public inquiries and national initiatives towards improved and more-visible regulatory oversight and monitoring. Large numbers of performance indicators are becoming publically available, leading to inevitable aggregation to produce rankings and league tables. But how reliable and useful are these?

Previous academic work by statisticians suggests that extremes of performance ("outliers") may be identifiable, which is valuable as these are of particular concern to all stakeholder groups, but that between the extremes little discrimination is possible – so league tables are dubious. These studies have used statistical techniques to analyse the impact of uncertainty (noise) in the underlying indicator data. These make parametric assumptions, which are questionable in many circumstances, and the impact of uncertainty about the aggregation weights has not been investigated beyond a few deterministic scenarios.

There is a powerful contribution that OR can make in this consequential area to go beyond conventional statistical approaches, and which is in line with the current OR agenda to engage with "data analytics". In this article, we use a large two-level Monte Carlo simulation to investigate the impact of varying the aggregation weights in a large set of indicators. With this assumption-light approach we demonstrate that identification of divergent performance is also possible in the presence of aggregation weight uncertainty.

The paper also promotes to OR analysts the use of the *R* software, which is free, scalable, conveniently handles high-dimensional data arrays, has a large library of statistical and other libraries and is very powerful for constructing customised graphical output.

Acknowledgement

The authors are members of a research team studying quality improvement capability in English NHS maternity services. The views expressed are those of the authors and not necessarily those of the Health Foundation.

Funding

The project has funding from the Health Foundation.

ORCID

Nathan C. Proudlove https://orcid.org/0000-0002-1176-8088
Mhorag Goff https://orcid.org/0000-0003-4936-2881
Kieran Walshe http://orcid.org/0000-0002-0696-480X
Ruth Boaden https://orcid.org/0000-0003-1927-6405

References

Ansoff, H. I. (1975). Managing strategic surprise by response to weak signals. *California Management Review 18*(2), 21–33.

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: Upcoming trends and challenges. *Journal of the Operational Research Society 60*(1), S16–S23.

Balci, O. (1998). Verification, validation, and testing. In J. Banks (Ed.), *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice* (pp. 335–393). Hoboken, NJ: John Wiley & Sons.

Banerjee, S., & Meitei, S. N. (2010). Effect of declining selling price: Profit analysis for a single period inventory model with stochastic demand and lead time. *Journal of the Operational Research Society 61*(4), 696–704.

Bardsley, M. (2016). Measuring and managing healthcare performance. In K. Walshe and J. Smith (Eds.), *Healthcare Management* (3rd ed., pp. 390–416). Maidenhead, UK: Open University Press.

Bardsley, M., Spiegelhalter, D. J., Blunt, I., Chitnis, X., Roberts, A., & Bharania, S. (2009). Using routine intelligence to target inspection of healthcare providers in England. *Quality and Safety in Health Care 18*(3), 189–194.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.

Bird, S., Cox, D. R., Farewell, V., Goldstein, H., Holt, T., & Smith, P. (2005). Working party on performance measures in the public services – Performance indicators: Good, bad, and ugly. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(1), 1–27.

Boaden. R., & Furnival, J. (2016). Quality improvement in healthcare. In K Walshe and J Smith (Eds.), *Healthcare*

Management (3rd ed., pp. 367–389). Maidenhead, UK: Open University Press.
Boulkedid, R., Alberti, C., & Sibony, O. (2013). Quality indicator development and implementation in maternity units. *Best Practice & Research Clinical Obstetrics & Gynaecology 27*(4), 609–619.
Brailsford, S. C., Harper, P., Patel, B., & Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation 3*(3), 130–140.
Burgess, F. J. (2012). Innovation and efficiency in health care: Does anyone really know what they mean? *Health Systems 1*(1), 7–12.
Care Quality Commission. (2015). *Newham University Hospital Quality Report 22/05/2015*. Retrieved from www.cqc.org.uk/sites/default/files/new_reports/AAAC0235.pdf
Carroll, F., Knight, H., Cromwell, D., Gurol-Urganci, I., & van der Meulen, J. (2016). *Patterns of maternity care in English NHS trusts 2013/14*. The Royal College of Obstetricians and Gynaecologists: 27 Sussex Place, Regent's Park, London NW1 4RG. Retrieved from www.rcog.org.uk/globalassets/documents/guidelines/research–audit/patterns-of-maternity-care-in-english-nhs-hospitals-2011-12_0.pdf
Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys 41*(3), 1–58.
Christensen, C. (2006). The ongoing process of building a theory of disruption. *Journal of Product Innovation Management 23*(1), 39–55.
Cumberlege, J. (2016). National Maternity Review. Better births: Improving outcomes of maternity services in England – A five year forward view for maternity care. NHS England. Retrieved from www.england.nhs.uk/wp-content/uploads/2016/02/national-maternity-review-report.pdf
Daraio, C., & Simar, L. (2016). Efficiency and benchmarking with directional distances: A data-driven approach. *Journal of the Operational Research Society 67*(7), 928–944.
Darley, S., Walshe, K., Boaden, R., Proudlove, N., & Goff, M. 2018. Improvement capability and performance: A qualitative study of maternity services providers in England. *International Journal for Quality in Health Care*. doi: 10.1093/intqhc/mzy081
Davey, M. (2015). *Review finds deaths of seven babies due to 'key failings' at Melbourne hospital Friday 16 October 2015*. The Guardian. Retrieved from www.theguardian.com/australia-news/2015/oct/16/review-finds-deaths-of-seven-babies-due-to-key-failings-at-melbourne-hospital
Department of Health. (1993). *Changing childbirth: Report of the Expert Maternity Group*. London: HMSO.
Department of Health. (2016). *Safer maternity care: Next steps towards the national maternity ambition*. Retrieved from www.gov.uk/government/publications/safer-maternity-care
Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring: A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society 61*(3), 486–501.
Flyvbjerg, B. (2011). Case study. In N. K. Denzin and Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (4th ed., pp. 301–316). Thousand Oaks, CA: SAGE.
Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations* (2nd ed.). New York: Wiley.
Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 159*(3), 385–443.
Graham, H., & Oakley, A. (1986). Competing ideologies of reproduction: Medical and maternal perspectives on pregnancy. In C. Currer and M. Stacey (Eds.), *Concepts of health, illness and disease: A comparative perspective*. Leamington Spa: Berg, pp. 99–115.
Griffiths, A., Beaussier, A.-L., Demeritt, D., & Rothstein, H. (2017). Intelligent monitoring? Assessing the ability of the care quality commission's statistical surveillance tool to predict quality and prioritise NHS hospital inspections. *BMJ Quality & Safety 26*(1), 120–130.
Harley, M., Mohammed, M. A., Hussain, S., Yates, J., & Almasri, A. (2005). Was Rodney Ledward a statistical outlier? Retrospective analysis using routine hospital data to identify gynaecologists' performance. *British Medical Journal 330*(7497), 929–933.
Hindle, A. G., Hindle, T., & Souli, S. (2009). Modelling and assessing local area differences in road casualties: A case study in England. *Journal of the Operational Research Society 60*(6), 781–788.
Hoad, K., Robinson, S., & Davies, R. (2010). Automating warm-up length estimation. *Journal of the Operational Research Society 61*(9), 1389–1403.
Howell, S. D., & Proudlove, N. C. (2007). A statistical investigation of inventory shrinkage in a large retail chain. *The International Review of Retail, Distribution and Consumer Research 17*(2), 101–120.
Jacobs, R., & Goddard, M. (2007). How do performance indicators add up? An examination of composite indicators in public services. *Public Money and Management 27*(2), 103–110.
Jacobs, R., Goddard, M., & Smith, P. C. (2005). How robust are hospital ranks based on composite performance measures? *Medical Care 43*(12), 1177–1184.
Karagiannis, G. (2017). On aggregate composite indicators. *Journal of the Operational Research Society 68*(7), 741–746.
Kennedy, I. (2001). *The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984–1995: Learning from Bristol*. Norwich: The Stationery Office. Retrieved from http://webarchive.nationalarchives.gov.uk/20090811143745/http:/www.bristol-inquiry.org.uk/final_report/the_report.pdf
Kirkup, B. (2015). *The report of the Morecambe Bay investigation*. Norwich: The Stationery Office. Retrieved from www.gov.uk/government/uploads/system/uploads/attachment_data/file/408480/47487_MBI_Accessible_v0.1.pdf
Klein, H. J., & Young, T. (2015). Health care: A case of hypercomplexity? *Health Systems 4*(2), 104–110.
Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(4), 835–851.
MacKenzie Bryers, H., & van Teijlingen, E. (2010). Risk, theory, social and medical models: A critical analysis of the concept of risk in maternity care. *Midwifery 26*(5), 488–496.
Macrae, C. (2014). Early warnings, weak signals and learning from healthcare disasters. *BMJ Quality & Safety 23*(6), 440–445.
Manktelow, B., Smith, L., Evans, T., Hyman-Taylor, P., Kurinczuk, J., Field, D., … Draper, E. (2015).

MBRRACE-UK perinatal mortality surveillance report: UK perinatal death for births from January to December 2013 – Supplementary report: UK Trusts and Health Boards. The Infant Mortality and Morbidity Studies Group, Department of Health Sciences, University of Leicester: Leicester. Retrieved from www.npeu.ox.ac.uk/mbrrace-uk/reports

Marshall, E. C., & Spiegelhalter, D. J. (1998). Reliability of league tables of in vitro fertilisation clinics: Retrospective analysis of live birth rates. *BMJ 316*(7146), 1701–1704.

Mingers, J. (2006). A critique of statistical modelling in management science from a critical realist perspective: Its role within multimethodology. *Journal of the Operational Research Society 57*(2), 202–219.

Mohammed, M. A., Cheng, K. K., Rouse, A., & Marshall, T. (2001). Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *The Lancet 357*(9254), 463–467.

Mohammed, M. A., Rathbone, A., Myers, P., Patel, D., Onions, H., & Stevens, A. (2004). An investigation into general practitioners associated with high patient mortality flagged up through the Shipman inquiry: Retrospective analysis of routine data. *British Medical Journal 328*(7454), 1474–1477.

Mohammed, M. A., Worthington, P., & Woodall, W. (2008). Plotting basic control charts: Tutorial notes for healthcare practitioners. *Quality & Safety in Health Care 17*(2), 137–145.

National Audit Office. (2013). *Maternity services in England.* Norwich, UK: The Stationary Office. Retrieved from www.nao.org.uk/wp-content/uploads/2013/11/10259-001-Maternity-Services-Book-1.pdf

Ord, K., & Fildes, R. (2013). *Principles of business forecasting*. Mason, OH: Cengage Learning.

Pidd, M. (2012). *Measuring the performance of public services: Principles and practice*. Cambridge, UK: Cambridge University Press.

Proudlove, N. C. (2012). Cracking the rankings Part (i): Understanding the Financial Times MBA rankings. *OR Insight 25*(4), 221–240.

Proudlove, N. C., Samarasinghe, B., & Walshe, K. (2018). Investigating consistent patterns of variation in short-notice cancellations of elective operations: The potential for learning and improvement through multi-site evaluations. *Health Services Management Research. 31*(3), 111–119.

Royston, G. (2013). Operational research for the real world: Big questions from a small island. *Journal of the Operational Research Society 64*(6), 793–804.

Savage, W. (2000). The caesarean section epidemic. *Journal of Obstetrics and Gynaecology 20*(3), 223–225.

Shaw, D., Westcombe, M., Hodgkin, J., & Montibeller, G. (2004). Problem structuring methods for large group interventions. *Journal of the Operational Research Society 55*(5), 453–463.

Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., & Grigg, O. (2012). Statistical methods for healthcare regulation: Rating, screening and surveillance; plus RSS discussion. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(1), 1–47. doi:10.1111/j.1467-985X.2011.01010.x

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine 24*(8), 1185–1202.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Talbot, C. (2010). *Theories of performance: Organizational and service improvement in the public domain*. Oxford: Oxford University Press.

Tennant, R., Mohammed, M. A., Coleman, J. J., & Martin, U. (2007). Monitoring patients using control charts: A systematic review. *International Journal for Quality in Health Care 19*(4), 187–194.

The King's Fund. (2008). *Safe births: Everybody's business.* London: King's Fund. Retrieved from www.kingsfund.org.uk/publications/safe-births-everybodys-business

The Royal College of Midwives. (2007). Making normal birth a reality: Consensus statement from the Maternity Care Working Party – Our shared views about the need to recognise, facilitate and audit normal birth, Maternity Care Working Party. Royal College of Midwives, Royal College of Obstetricians and Gynaecologists and National Childbirth Trust. Retrieved from www.rcm.org.uk/sites/default/files/NormalBirthConsensusStatement.pdf

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society 66*(8), 1352–1362.

Trapero, R. J., Kourentzes, N., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society 66*(2), 299–307.

Wallace, E. M. (2015). *Report of an investigation into perinatal outcomes at Djerriwarrh Health Services.* Melbourne, Australia: Victoria State Government. Retrieved from www2.health.vic.gov.au/hospitals-and-health-services/quality-safety-service/djerriwarrh

Walshe, K., Addicott, R., Boyd, A., Robertson, R., & Ross, S. (2014). *Evaluating the Care Quality Commission's acute hospital regulatory model: Final report.* Manchester Business School and The King's Fund. Retrieved from www.cqc.org.uk/sites/default/files/CM071406%20Item%206%20Acute%20regulatory%20model%20evaluation%20report.pdf

Wood, M., Capon, N., & Kaye, M. (1998). User-friendly statistical concepts for process monitoring. *Journal of the Operational Research Society 49*(9), 976–985.

Yan, X., Bai, Y., Fang, S.-C., & Luo, J. (2016). A kernel-free quadratic surface support vector machine for semi-supervised learning. *Journal of the Operational Research Society 67*(7), 1001–1011.

Zhao, Y., Zhao, X., & Shen, Z.-J. M.. (2017). On learning process of a newsvendor with censored demand information. *Journal of the Operational Research Society 67*(9), 1200–1211.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: The University of Michigan Press.