Using quality indicators to predict inspection ratings:
cross-sectional study of general practices in England

INTRODUCTION
High-quality general practice is an essential component of a well-functioning healthcare system, but performance varies across practices. One approach that is increasingly used in an attempt to maintain and improve high standards of care is regulation, and most healthcare regulators across the world make use of some form of provider inspections. Typically, these are pre-announced visits that follow a set structure and evaluate providers against certain standards or requirements. Depending on the size of the provider and the intensity of the inspection, these visits can require significant amounts of time and human resources, and they are an expensive method of assessing quality. As a result of this high cost, it has been suggested that inspections should, where possible, be targeted where the potential for improvement is greatest and most needed.

In England, health and social care providers are regulated by the Care Quality Commission (CQC), which was established in 2009 when the two previous regulators of health care and social care were merged. In the subsequent couple of years, several failures in care and management became very prominent in the media and politically. These included:

- excess deaths and poor care at Stafford Hospital, an acute hospital run by Mid Staffordshire NHS Foundation Trust — in June 2010, the UK government announced there would be a public inquiry;
- patient abuse by staff at Winterbourne View private residential hospital for people with learning disabilities — this was exposed by a BBC television documentary in May 2011; and
- the July 2011 financial collapse of Southern Cross Healthcare Group, a private provider of health and social care services — at that time, it was the largest provider, with 31 000 residents in 750 care homes.

These were noted in criticism of the CQC by the National Audit Office, the House of Commons Health Committee, and the Department of Health’s performance and capability review. It was also highlighted that between October 2010 and March 2011, the CQC had completed fewer than half of its planned number of inspections, because of prioritisation of registration over compliance and a shortage of inspectors. The CQC responded by introducing a new, comprehensive inspection regime intended to help it ‘make better decisions about when, where, and what to inspect by using information and evidence in a more focused and open way’. Part of this regime included developing a system, known as Intelligent Monitoring (IM), using performance indicators based on routinely collected and available data about providers. The IM system for general practices contained 33 performance indicators concerning
patient satisfaction, the management of chronic conditions, prescribing, disease prevalence, and emergency hospital admissions. Together, these indicators were intended to support inspections with information on which practices to inspect and on which aspects of care to focus on.\textsuperscript{12,13} Performance indicators were aggregated into ratings of potential risk based on expected indicator values; these were to be used to create priority bands for inspection.\textsuperscript{13} Between September 2014 and January 2017, the CQC inspected 7330 practices, and awarded each of them one of four rating levels (‘outstanding’, ‘good’, ‘requires improvement’, or ‘inadequate’) in five domains (‘safe’, ‘effective’, ‘responsive’, ‘caring’, and ‘well led’), along with an ‘overall’ rating that summarised the domain ratings.

Although the CQC’s IM indicators were not intended to replace inspections, this study examined whether they can be used to predict subsequent practice inspection ratings. This is important in terms of whether IM indicators can legitimately fulfil their aim to prioritise practice inspections or help determine the focus of inspection visits.

The equivalent IM system for secondary care has been found to be of limited value in prioritising hospital inspection or determining the focus of inspection visits.\textsuperscript{14,15} Recent research has examined the CQC inspection process and its impact on performance. Regarding the process, CQC inspectors have been found to disagree on judgements on vignettes of real inspection data, so assessments by teams are considered more reliable,\textsuperscript{16} and inspections were found to involve service users’ views, but in a transactional manner and on the CQC’s terms.\textsuperscript{17} Regarding their impact, IM indicators for acute hospital trusts have been found to lack predictive power\textsuperscript{18} and inspection of A&E departments has been found to have no impact on their performance metrics.\textsuperscript{19}

How this fits in
Recent studies have questioned the utility of quality indicators in monitoring acute-hospital provider performance and targeting regulatory inspections; this study now extends this to general practice. It was found that the quality indicators in use from 2014 until 2017 were of little value in predicting subsequent inspection ratings and, as a result, suggestions regarding the future development of indicators include using more recent and up-to-date data, drawn from a wider range of sources, and considering changes over time.

METHOD
Data
The published inspection ratings and corresponding inspection dates for the 7330 GP practices that were inspected between September 2014 and January 2017 were obtained on request from the CQC. Over the inspection cycle, some practices were re-inspected and their ratings updated. Only the rating from the first inspection was used in this study, as subsequent ratings were likely to have been influenced by the previous inspection process and outcomes. Data on the most recent practice inspections are freely available online.\textsuperscript{19}

For each of the inspected practices, the following data were obtained:

- the date of inspection;
- the date the rating was made public;
- the ‘overall’ rating; and
- the rating in each domain.

The IM dataset was also obtained from the CQC. There were two releases of IM, with indicator values being updated as new data were published. The second release\textsuperscript{13} was used; this covered the time period from April 2013 until December 2014 and, therefore, corresponded to the majority (90\%) of inspections. There were no further updates to IM, so the indicator values for practices inspected later in the cycle were more than 2 years old.

The indicators were selected by the CQC through consultation and testing, and detailed definitions (Box 1) and data sources were made available.\textsuperscript{13} Data on one of the indicators — relating to emergency cancer admissions (GPHLLIEC01) — were censored by the CQC for 1200 practices because of small numbers of admissions, resulting in a missing indicator value for these practices. As the statistical methods required a complete set of indicators, it was decided that this frequently missing indicator would be removed from the analysis. Each indicator was categorised by the CQC as relating to the ‘effective’, ‘caring’, or ‘responsive’ key domain; there were no indicators categorised to the other two domains (‘safe’ and ‘well led’) (Box 1). The raw indicator values were used to fully take account of the variations in indicators across practices.

The inspection ratings and IM datasets were merged to form one dataset containing rating and indicator data for the 6860 practices both monitored and inspected (470 practices were not included in the monitoring data).
### Box 1. The 33 Intelligent Monitoring indicators for general practice

| Domain       | Indicator code | Indicator description                                                                 |
|--------------|----------------|---------------------------------------------------------------------------------------|
| Effective    | GPHLIA01       | Number of emergency admissions for 19 ambulatory care-sensitive conditions per 1000 population |
| Effective    | GPHLIF01       | Number of antibacterial prescription items prescribed per STAR PU                      |
| Effective    | GPHLIF02       | Ratio of reported versus expected prevalence for coronary heart disease                  |
| Effective    | GPHLIF03       | Ratio of reported versus expected prevalence for chronic obstructive pulmonary disease |
| Effective    | GPHLIF04       | Percentage of antibiotic items prescribed that are cephalosporins or quinolones        |
| Effective    | GPHLIF05       | Emergency cancer admissions per 100 patients on disease register                         |
| Effective    | GPHLIF06       | Percentage of patients aged ≥65 years who have received a seasonal influenza vaccination |
| Effective    | GPHLIF07       | Percentage of patients aged ≥65 years who have received a seasonal flu vaccination       |
| Effective    | GPHLIF08       | Percentage of patients aged ≥65 years whose last measured total cholesterol was ≤140/80 mmHg |
| Effective    | GPHLIF09       | Percentage of patients aged ≥65 years whose last measured albumin:creatinine ratio test was ≤64 mmol/mol |
| Effective    | GPHLIF10       | Percentage of patients with diabetes, on the register, whose last blood pressure reading in the preceding 12 months was ≤140/80 mmHg |
| Effective    | GPHLIF11       | Percentage of patients with diabetes, on the register, whose last measured total cholesterol was ≤150/90 mmHg |
| Effective    | GPHLIF12       | Percentage of patients with hypertension in whom the last blood pressure reading measured in the preceding 9 months was ≤150/90 mmHg |
| Effective    | GPHLIF13       | Percentage of patients with physical and/or mental health conditions whose notes recorded smoking status in the preceding 12 months |
| Effective    | GPHLIF14       | Percentage of patients aged ≥75 years with a fragility fracture on or after 1 April 2012, who are currently treated with an appropriate bone-sparing agent |
| Effective    | GPHLIF15       | Percentage of patients diagnosed with dementia whose care has been reviewed in a face-to-face review in the preceding 12 months |
| Effective    | GPHLIF16       | Percentage of patients with diabetes, on the register, who have had a record of an albumin:creatinine ratio test in the preceding 12 months |
| Effective    | GPHLIF17       | Percentage of patients with diabetes, on the register, whose last measured total cholesterol (measured in the preceding 12 months) was ≤5 mmol/L |
| Effective    | GPHLIF18       | Percentage of patients with diabetes, on the register, who have had influenza immunisation in the preceding 1 September to 31 March |

*The GP Patient Survey is an independent survey run by Ipsos MORI on behalf of NHS England. The survey is sent out to over a million people across the UK. The results show how people feel about their GP practice. CHADS2 = Congestive heart failure, Hypertension, Age ≥75, Diabetes mellitus, and prior Stroke or transient ischaemic attack. IFCC-HbA1c = International Federation of Clinical Chemistry glycated haemoglobin (average plasma glucose concentration) measure. STAR PU = Specific Therapeutic group Age-sex Related Prescribing Unit.*
Statistical analysis
The rating levels were transformed into an ordered categorical variable ranging from 1 ('inadequate') to 4 ('outstanding'). Ordered logistic regression was used to model the relationship between the 'overall' ratings and the 32 IM indicators. The same methods and modelling approach were applied to the three domain ratings on their domain-specific indicators. From these regression results, the probability that a practice received each of the four rating scores was predicted. These probabilities were multiplied by the rating score (1–4) and summed to arrive at a single, expected predicted value. This continuous predicted value was transformed to a discrete predicted rating, based on the following rules:
- predicted rating = 1 ('inadequate') if predicted values \( \leq 1.5 \);
- predicted rating = 2 ('requires improvement') if predicted values were \( > 1.5 \) and \( \leq 2.5 \);
- predicted rating = 3 ('good') if predicted values were \( > 2.5 \) and \( \leq 3.5 \); and
- predicted rating = 4 ('outstanding') if predicted values were \( > 3.5 \).

This approach and set of rules approximated a simple prediction model that the CQC could have used when prioritising inspections based on existing information.

After fitting the regression model and generating the predicted ratings, a measure of agreement (weighted \( \kappa \)) was used to compare the predicted with the actual ratings, in line with methodologies outlined by Cohen\(^2\) and Jakobsson and Westergren.\(^2\) Cohen's weighted \( \kappa \) accounts for the level of agreement that would be expected by chance, and the test statistic measures...
the degree to which the predicted ratings improve — or not — on this level of chance agreement. Weighted $\kappa$ ranges from -1 to 1, with a value close to 0 indicating little or no agreement beyond that expected through chance alone; a value close to 1 indicates strong agreement. A negative value indicates a level of agreement that is less than that expected by chance — that is, disagreement — while values in the 0.6–0.8 range indicate good agreement.21 The usual weighting system was applied, as follows:

- when predicted and observed ratings were equal, the weighting was 1;
- when ratings differed by one or two rating levels, the weighting was two-thirds, or one-third respectively; and
- when the difference was at the maximum (three rating levels), the weighting was 0.20.21

In an alternative approach, the four-level rating score was dichotomised by grouping ‘inadequate’ with ‘requires improvement’, and ‘good’ with ‘outstanding’. The probability of being in either of these two groups was estimated by logistic regression.

**RESULTS**

Figure 1 plots the inspections for each month dating from September 2014 until January 2017 for all 6860 practices in the dataset. The majority (5406 out of 6860, 78.8%) of practices were rated ‘good’; very few (275 out of 6860, 4.0%) were rated ‘inadequate’. A greater number of practices were inspected later in the cycle (from 2016 onwards) compared with earlier on. If the CQC had been using the IM dataset to prioritise inspections of practices that appeared to perform less well on the IM indicators, it could be expected that there would have been a pattern of more ‘inadequate’ and ‘requires improvement’ ratings earlier in the cycle; however, such a pattern was not apparent.

Figure 2 plots the percentage of practices that received each rating for each of the five domains and for the ‘overall’ rating. The ratings were not particularly discriminating: in all domains, the majority of practices were rated ‘good’. The ‘caring’ domain had the highest proportion of ‘good’ ratings and the ‘safe’ domain had the lowest.

**Regression results**

There were complete data on all 32 IM indicators for 5988 practices (87.3% of the dataset); these were used to predict the ‘overall’ rating. At the domain level, 6567 practices (95.7%) had complete data for the 24 indicators in the ‘effective’ domain, 6149 (89.6%) had complete data for all six indicators in the ‘caring’ domain, and 6835 (99.6%) for the two indicators in the ‘responsive’ domain.

Table 1 compares the predicted ‘overall’ ratings from the ordered logistic regression of ‘overall’ practice rating and all Intelligent Monitoring indicators with actual ‘overall’ inspection ratings, and uses the percentage of correct predictions and weighted $\kappa$ to assess the level of agreement. The regression model is shown in Table 2. It was found that 79.7% of actual ratings were predicted correctly; however, this occurred because the model predicted the great majority of practices to be rated ‘good’. Specifically, there were 4818 practices rated ‘good’ after inspection; 5756 practices had a predicted rating of good.
that none of the 270 practices that were deemed ‘outstanding’ would be awarded such a rating (0.0% accuracy).

The weighted κ statistic of 0.097 meant that predicted ratings were 9.7% of the way between the level of agreement expected by chance and perfect agreement — this is very low, and far below the 0.6–0.8 range that would be considered good agreement.20,21 Figure 3 illustrates this point graphically by plotting the predicted inspection rating scores (prior to rounding to discrete categories) against actual inspection ratings. There was substantial overlap between the ranges of predicted inspection ratings for each of the four sets of actual inspection rating outcomes. Comparable models and predictions for the ‘caring’, ‘effective’, and ‘responsive’ domains are not shown but performed worse, with fewer accurate predictions and lower weighted κ statistics.

The dichotomised ratings score was used in a binary logistic regression to predict the probability of receiving a poor inspection rating (‘inadequate’ or ‘requires improvement’) compared with a good inspection rating (‘good’ or ‘outstanding’), based on IM indicator values. The predicted probabilities for both groups indicated a higher degree of overlap than shown in Figure 3 and it was not possible to distinguish practices in greater need of inspection.

**DISCUSSION**

**Summary**

The performance indicators in the CQC’s IM dataset were found to have very limited ability to predict the inspection ratings of general practices in advance of inspection, and they also could not separate practices as being merely good or bad. Moreover, predictions were more likely to underestimate the need for inspection by predicting practices to have a better rating than was found on inspection. A data-driven approach to the prioritisation of general practice inspections using the IM indicators is unlikely to be effective unless significant improvements are made.

**Strengths and limitations**

This study used all but one of the IM indicators (emergency cancer admissions), and their raw values, examination of all domains with allocated indicators, and the large dataset of GP practices. Had the models found a reasonable fit to this entire dataset, split-sample testing or cross-validation could have been conducted to get a fairer estimation of the predictive
accuracy by fitting a model to a subset of GP practices chosen randomly or by inspection date [temporal validation], and then applying the model to obtain predictions for the remainder of practices. The predictive ability of such out-of-sample testing would likely be lower than that found in this study’s whole-dataset results because the model would have been tested on a subset of data different from that to which it had been fitted (tuned).

One limitation was inherent in the inspection dataset: as actual inspection ratings were so undiscriminating, and were asymmetrically distributed across the four categories with nearly 80% of ‘overall’ ratings being ‘good’, it would have been difficult for any model to achieve predictive accuracy better than chance. Furthermore, it was only possible to test the power of IM at predicting domain ratings when the IM indicators were targeted at those domains (‘effective’, ‘caring’, and ‘responsive’). It was not possible to do this for the domains of ‘well led’ and ‘safe’, as these had no IM indicators; practices tended to have poorer ratings in these domains.

Comparison with existing literature
A similar lack of predictive ability was found when analysing the IM system for acute hospitals, both overall, and on the underlying domains [domains such as ‘safe’ and ‘caring’]. There have not been any studies that find value in the IM data.

Implications for research and practice
It should be noted that CQC guidance on IM says that it was used to ‘raise questions, not make judgements’, and that inspections draw on a much wider range of both quantitative and qualitative data; as such, one would not necessarily expect to find a very high level of agreement between predicted and actual inspection ratings. However, as so little agreement was found, it has brought into question how useful the IM indicators are to the inspection programme — it could also throw doubt on the validity and reliability of CQC inspection ratings, the IM indicators, or both, as measures of practice quality. The findings presented here raise important questions about what those indicators and the CQC inspection ratings were supposed to be measuring. If both sought to measure the quality of care or wider aspects of general practice performance, it is perhaps a surprise and a cause for concern that there seems to be so little agreement between them. Of course, the fact that CQC ratings for general practices were so unevenly distributed made it statistically more difficult for any dataset to have predictive value.

In its strategy for 2016–2021, the CQC has articulated its aim to move to a more intelligence-led model of regulation and inspection across all health and social care sectors, along with its plans to replace IM with a new system named CQC Insight. Given that, it is worth considering the lessons that can be drawn from the research presented here. It is always important to be clear about the purpose of routine monitoring of performance indicators, and to incorporate testing and validation into their development and piloting before they are deployed in routine
It may be attractive to make use of existing and available datasets, but their validity as measures of the quality of care or wider performance should be explored and tested. Such datasets will never be capable of predicting the outcomes of inspection perfectly (and, of course, if they were, there would be no need for such inspections), because inspections are able to draw on a much wider range of qualitative and quantitative data, as well as on the expertise and judgement of the inspection team. However, one should expect IM, CQC Insight, and other forms of routine monitoring to have some predictive value. It may be possible to improve their predictive value by using:

- more recent and up-to-date data;
- time-series data to take into account changes in provider performance over time; and
- a wider range of data from other sources, such as patient feedback.24

In conclusion, it seems that CQC and other regulators should test and evaluate the validity and reliability of both performance indicator systems and inspection rating schemes if and when they are introduced; and arguably should continue to monitor validity and reliability routinely. If their predictive value cannot be demonstrated, this will limit the scope for CQC and other regulators to use risk-based or responsive regulatory models that target regulatory interventions such as provider inspections based on their reported performance.
REFERENCES

1. Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the Quality and Outcomes Framework. Lancet 2008; 372(9640): 728–736.

2. Doran T, Campbell S, Fullwood C, et al. Performance of small general practices under the UK’s Quality and Outcomes Framework. Br J Gen Pract 2010; DOI: https://doi.org/10.3399/bjgp10X51340.

3. Evans E, Akings H, Edwards A. Reducing variation in general practitioner referral rates through clinical engagement and peer review of referrals: a service improvement project. Qual Prim Care 2011; 19(4): 263–272.

4. Department of Health. Tackling health inequalities — a programme for action. 2013. https://webarchive.nationalarchives.gov.uk/20031221042257/http://www.doh.gov.uk/healthinequalities/programmeforaction/ (accessed 5 Dec 2019).

5. Gillam S, Srinivasaena AN. Regulation in primary care. Qual Prim Care 2014; 22(2): 57–61.

6. Lewis R, Alvarez-Rosete A, Mays N. How to regulate health care in England? An international perspective. 2006. https://www.kingsfund.org.uk/sites/default/files/fileld/publication_file/how-regulate-health-care-england-international-perspective-richard-lewis-arturo-alvarez-rosete-nicholas-mays-kings-fund-23-november-2006.pdf (accessed 5 Dec 2019).

7. Ayres I, Braithwaite J. Oxford Socio-Legal Studies. Responsive regulation: transcending the deregulation debate. Oxford: Oxford University Press, 1992.

8. National Audit Office. The Care Quality Commission: regulating the quality and safety of health and adult social care. 2011. http://www.nao.org.uk/wp-content/uploads/2011/12/10121665.pdf (accessed 5 Dec 2019).

9. House of Commons Health Committee. 2012 Accountability hearing with the Care Quality Commission: Seventh Report of Session 2012–13. 2012. http://www.publications.parliament.uk/pa/cm201213/cmselect/cmhealth/592/592.pdf (accessed 5 Dec 2019).

10. Department of Health. Performance and capability review: Care Quality Commission. 2012. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/215158/dh_123791.pdf (accessed 5 Dec 2019).

11. Care Quality Commission. A new start: consultation on changes to the way CQC regulates, inspects and monitors care. 2013. http://www.cqc.org.uk/sites/default/files/documents/cqc_consultation_2013_tagged_0.pdf (accessed 5 Dec 2019).

12. Care Quality Commission. Raising standards, putting people first: our strategy for 2013 to 2016. 2013. https://www.cqc.org.uk/sites/default/files/documents/20130503_cqc_strategy_2013_final_cm_tagged.pdf (accessed 5 Dec 2019).

13. Care Quality Commission. Intelligent Monitoring: NHS GP practices. Indicators and methodology. 2014. http://www.cqc.org.uk/sites/default/files/20141205GPIMIndicatorsAndMethodologyGuidance.pdf (accessed 5 Dec 2019).

14. Griffiths A, Beausssier AL, Demeter D, Rothstein H. Intelligent Monitoring? Assessing the ability of the Care Quality Commission’s statistical surveillance tool to predict quality and prioritise NHS hospital inspections. BMJ Qual Saf 2017; 26(2): 120–130.

15. Allen T, Walshe K, Proudlove N, Sutton M. Do performance indicators predict regulator ratings of healthcare providers? Cross-sectional study of acute hospitals in England. Int J Qual Health Care 2019; DOI: https://doi.org/10.1093/ijqhc/mrz101.

16. Boyd A, Addicott R, Robertson R, et al. Are inspectors’ assessments reliable? Ratings of NHS acute hospital trust services in England. J Health Serv Res Policy 2017; 22(1): 28–36.

17. Richardson E, Walshe K, Boyd A, et al. User involvement in regulation: a qualitative study of service user involvement in Care Quality Commission inspections of health and social care providers in England. Health Expect 2019; 22(2): 245–253.

18. Allen T, Walshe K, Proudlove N, Sutton M. Measurement and improvement of emergency department performance through inspection and rating: an observational study of emergency departments in acute hospitals in England. Emerg Med J 2019; 36(6): 326–332.

19. Care Quality Commission. Using CQC data. 2019. https://www.cqc.org.uk/about-us/transparency/using-cqc-data (accessed 5 Dec 2019).

20. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968; 70(4): 213–220.

21. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci 2005; 19(4): 427–431.

22. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009; 338: b604.

23. Care Quality Commission. Shaping the future: CQC strategy for 2016 to 2021. 2016. http://www.cqc.org.uk/sites/default/files/20160523_strategy_16-21_strategy_final_web_01.pdf (accessed 5 Dec 2019).

24. Griffiths A, Leaver MP. Wisdom of patients: predicting the quality of care using aggregated patient feedback. BMJ Qual Saf 2018; 27(2): 110–118.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Allen, T; Walshe, K; Proudlove, N; Sutton, M

Title:
Using quality indicators to predict inspection ratings: cross-sectional study of general practices in England

Date:
2020-01-20

Citation:
Allen, T., Walshe, K., Proudlove, N. & Sutton, M. (2020). Using quality indicators to predict inspection ratings: cross-sectional study of general practices in England. British Journal of General Practice, 70 (690), pp.e55-e63. https://doi.org/10.3399/bjgp19x707141.

Persistent Link:
http://hdl.handle.net/11343/247225

File Description:
published version

License:
CC BY