Understanding high and low patient experience scores in primary care: analysis of patients’ survey data for general practices and individual doctors

Martin J Roberts research fellow¹, John L Campbell professor of general practice and primary care¹, Gary A Abel senior research associate², Antoinette F Davey research fellow¹, Natasha L Elmore research assistant³, Inocencio Maramba associate research fellow¹, Mary Carter associate research fellow¹, Marc N Elliott senior principal researcher, chair in statistics³, Martin O Roland professor of health services research², Jenni A Burt research associate²

¹University of Exeter Medical School, St Lukes Campus, Exeter EX1 2LU, UK; ²Cambridge Centre for Health Services Research, Institute of Public Health, Forvie Site, University of Cambridge School of Clinical Medicine, Box 113 Cambridge Biomedical Campus, Cambridge CB2 0SR, UK; ³RAND Corporation, 1776 Main Street, PO Box 2138, Santa Monica, CA 90407-2138, USA

Abstract

Objectives To determine the extent to which practice level scores mask variation in individual performance between doctors within a practice.

Design Analysis of postal survey of patients’ experience of face-to-face consultations with individual general practitioners in a stratified quota sample of primary care practices.

Setting Twenty-five English general practices, selected to include a range of practices scores on doctor-patient communication items in the English national GP Patient Survey.

Participants 7721 of 15172 patients (response rate 50.9%) who consulted with 105 general practitioners in 25 practices between October 2011 and June 2013.

Main outcome measure Score on doctor-patient communication items from post-consultation surveys of patients for each participating general practitioner. The amount of variance in each of six outcomes that was attributable to the practices, to the doctors, and to the patients and other residual sources of variation was calculated using hierarchical linear models.

Results After control for differences in patients’ age, sex, ethnicity, and health status, the proportion of variance in communication scores that was due to differences between doctors (6.4%) was considerably more than that due to practices (1.8%). The findings also suggest that higher performing practices usually contain only higher performing doctors. However, lower performing practices may contain doctors with a wide range of communication scores.

Conclusions Aggregating patients’ ratings of doctors’ communication skills at practice level can mask considerable variation in the performance of individual doctors, particularly in lower performing practices. Practice level surveys may be better used to “screen” for concerns about performance that require an individual level survey. Higher scoring practices are unlikely to include lower scoring doctors. However, lower scoring practices require further investigation at the level of the individual doctor to distinguish higher and lower scoring general practitioners.

Introduction

Public reporting of performance measures is increasingly the norm in healthcare systems.¹ Forming part of the drive for continuous quality improvement, the disclosure of results of assessments at either provider or individual level is believed to increase accountability and public engagement.² A recent US report highlighted the important contribution that listening to, and acting on, patients’ feedback can potentially make to efforts to improve healthcare.³ New developments in the English National Health Service highlight the embedding of public assessment of performance within the regulation of the healthcare system, including NHS England’s consultation on the production of general practice league tables and the Care Quality Commission’s parallel development of a rating system for primary care.⁴ ⁵ An increasingly transparent healthcare system, in which providers are publicly gauged against performance targets, is regarded by policy makers as essential
to enabling patients to make informed choices about the care they receive. Consequently, patients’ feedback on healthcare services is now gathered in the United States, Canada, Europe, Australia, China, and elsewhere.

This increased emphasis on patients’ feedback in healthcare is reflected in extensive investment in both collection and use of patients’ experience data to evaluate providers’ performance. In the United Kingdom, for example, the NHS Outcomes Framework 2013/14 requires that “people have a positive experience of care.” For primary care, this is assessed on the basis of responses to the English national GP Patient Survey of patients’ experiences with their general practitioner surgery. This major source of patient experience data, currently administered to over 2 million people annually, is also the source for general practices’ performance scores compiled for and advertised on websites such as NHS Choices and Compare.

Similar internet based formats for reporting patient experience data, whether generated by governments, patient groups, or commercial organisations, are emerging across the globe.

Several causal pathways for achieving improvements in providers’ performance through the release of publicly reported performance data have been proposed. Some invoke market-like selection, claiming that patients will modify their choice of provider by using publicly available data, such as that provided by patient experience websites. Evidence to support this pathway is, however, weak. A more likely mechanism driving improvement in performance in response to the publication of performance data is health professionals’ concern for reputation, in which peer comparison motivates individuals and organisations to improve their care.

Irrespective of its potential to stimulate change, the publication of performance data is central to the openness and transparency that are seen as essential to a safe, equitable, patient centred healthcare system. Thus, regardless of any effect on quality improvement, such initiatives are likely to be here to stay. In refining the information made public, performance data need to be accurate and relevant to all potential users. The US based Robert Wood Johnson Foundation has noted that although patients “prefer to see comparative information for individual providers rather than practices or groups,” this is often not done in practice. Currently, however, the focus is moving from the publication of performance data at an organisational level to that of individual doctors. In the United Kingdom, for example, patients referred to the cardiology service at the South Manchester Hospitals Trust may go online to view both mortality and patient experience data for each cardiologist or cardiac surgeon. However, in English primary care, nationally collected performance data remains at the level of the practice, not the practitioner. The practice level aggregation of data from the GP Patient Survey, used to derive practices’ performance indicators, potentially masks considerable variation in performance among individual general practitioners, thereby inappropriately advantaging or disadvantaging particular doctors.

Current indicators may consequently fail to provide users, providers, or commissioners with an accurate assessment of performance within a practice.

We aimed to explore the extent to which aggregated practice scores may mask variation in individual performance. We focused on patients’ assessments of doctors’ communication skills: interpersonal aspects of care are a key driver of overall patients’ satisfaction and are a major component of the GP Patient Survey used to derive the “overall patient experience of care” scores advertised for each general practice on NHS Choices (alongside waiting to be seen, opening hours, and consultations with a nurse).

**Methods**

**Sample and data collection**

We invited a stratified random sample of general practices in six areas of England (Cornwall, Devon, Bristol, Bedfordshire, Cambridgeshire, and North London) to participate in the study. Our aim was to recruit a set of practices that varied substantially in ratings for doctors’ consultation skills. Practices were classified at baseline according to their case mix adjusted score on a composite of seven items relating to doctor-patient communication in the 2009/10 English GP Patient Survey. We used linear regression models adjusting for patients’ age, sex, ethnicity, deprivation score, and self rated health to adjust for case mix. We aimed to recruit 15 practices with scores in the lowest 25% of all practices, five practices scoring in the middle quarter (37.5th to 62.5th centile), and five practices scoring in the highest quarter (above the 75th centile). Eligible practices had at least two registered general practitioners working at least four sessions a week (0.4 full time equivalent), excluding trainees and short term locums. We drew a stratified random sample, stratifying by the communication score banding, general practitioner head count, deprivation index, and geographical location. We approached eligible practices in a randomised order until the quota for each stratum was obtained.

Data collection took place between October 2011 and June 2013. We did a postal survey of patients who had recently attended a face-to-face consultation with a participating general practitioner. For each wave of the survey, we extracted from electronic records a list of face-to-face doctor-patient consultations held during the previous three weeks. Practices screened each list for recent deaths, recent bereavement, terminal illness, and mental incapacity: all such patients were excluded. Practices sent the remaining patients a patient experience survey based on the national GP Patient Survey, asking them about access, waiting times, opening hours, and continuity and interpersonal aspects of care. The questionnaire also included questions about sociodemographic information including age, sex, ethnicity, and self rated health. In completing the seven interpersonal care items and one confidence and trust item, patients were asked to think back to a consultation with a specified doctor on a specified date (corresponding to the consultation identified from the extracted records). Patients who attended multiple consultations were sent only one questionnaire, relating to their most recent consultation at the point of data extraction. One reminder was sent to patients who did not respond within three weeks; we accepted returned questionnaires up to 100 days after the initial mail out. We repeated the survey cycle in each practice until either we had received 50 or more completed questionnaires for each participating general practitioner or three cycles had been completed. Fifty questionnaires are sufficient to obtain reliable mean communication scores for comparable patient feedback instruments. Return of a completed questionnaire was taken to indicate patients’ consent to participate in the study.

**Outcome measures**

Our primary outcome measure was a communication score for the doctor from each respondent. We derived this as the mean rating across the seven communication items (questions 22a to 22g; see supplementary material) among patients providing four or more informative responses. The first of our five secondary outcomes asked patients about their confidence and trust in the doctor they saw. We also analysed four secondary measures relating to practice level variables: patients’ ratings of overall satisfaction with care at the practice, helpfulness of the

[98x303]16

[104x741]6

[117x237]17

[117x237]18

[271x551]11-13

[272x369]2

[278x584]8-10

[307x35]satisfaction with care at the practice, helpfulness of the

[307x46]practitioner. Our primary outcome measure was a communication score for the doctor from each respondent. We derived this as the mean rating across the seven communication items (questions 22a to 22g; see supplementary material) among patients providing four or more informative responses. The first of our five secondary outcomes asked patients about their confidence and trust in the doctor they saw. We also analysed four secondary measures relating to practice level variables: patients’ ratings of overall satisfaction with care at the practice, helpfulness of the
receptionists, cleanliness of the facilities, and ease of entry to
the building. All items were rescaled linearly from 0 (least
favourable) to 100 (most favourable).

**Statistical analysis**

We described the sex balance, proportion of doctors who trained
in the United Kingdom, and mean time since registration in the
general practitioner sample, together with questionnaire response
rates and the intervals between consultations with patients, mail
out of questionnaires, and return of questionnaires. We tested
whether consultation to mail out intervals were associated with
response to questionnaires by using a two sample \(t\) test.

In our study design, groups of individual patients’ scores are
associated with (nested within) individual general practitioners,
and groups of general practitioners are associated with individual
practices. Although some variance in patients’ scores can be
attributed to individual experiences (including personal
expectation, outlook, and the variable performance of general
practitioners between patients), some of the variance in patients’
scores is likely to be attributable to general practitioners, with
some doctors performing better, on average, than others.

Furthermore, aspects of the practices beyond the general
practitioners (for example, reception staff, opening hours) may
account for some variation. Our aim was to assess the extent to
which aggregated practice scores may mask within practice
variation in performance by individual general practitioners.

We used three-level mixed-effects hierarchical linear models
to estimate the amount of variance in each of the outcome
measures that could be attributed to differences between the
practices, to differences between the doctors within each
practice, and to the patients and other residual sources.22 Such
models represent an extension of analysis of variance based
approaches, taking account of the inherent hierarchical structure
of the data: patients are clustered within doctors, who in turn
are clustered within practices.

We adjusted all models for four self reported patients’ attributes
previously shown to be important predictors of reported patient’s
experience: the patient’s sex, age (eight ordinal categories),
ethnicity (16 categories), and self reported health status (five
ordinal categories).21 We expressed the practice, doctor, and
patient related variance components from each model as
percentages of the total variance and used the “best linear
unbiased predictors” of the practice and doctor effects to provide
estimates of the mean score for each doctor on each of the
outcome measures.23 Corresponding estimates of the mean scores
for each practice came from additional models, which omitted
random effects for doctors. We described the variation in the
general practitioners’ and practices’ mean scores and used
simple correlation analysis to investigate the association between
the practices’ mean score and the within practice standard
deviation of the general practitioners’ mean scores. We used
the variance components from each model to estimate the
number of patients’ scores per doctor needed to achieve a
reliability of at least 0.7 or 0.8 for the doctor’s mean score (see
appendix for the formula used). Whereas a reliability of 0.8 or
higher is desirable for moderate to high stakes assessments,24 a
threshold of 0.7 is regarded as acceptable in patients’
assessments of doctors’ performance in some contexts.25 We
used Stata SE version 10.1 for data analysis.

**Results**

Of 59 practices initially approached, six were found to be
ineligible, nine declined participation, and 19 had not responded
by the time we achieved our quota of 25 participating practices.

Table 1 provides brief profiles of the participating practices.
There were 105 participating doctors (mean 4.2 (range 2-8)
doctors per practice), of whom 46% were female and 80% were
trained in the United Kingdom. Average time since registration
with the General Medical Council was 19.5 (range 4-38) years.
Table 2 shows respondents’ demographics. The mean interval
between the patient’s consultation date and the mail out of their
questionnaire was 16.6 (SD 6.0) days. We found no evidence
that the length of this interval was related to the likelihood of
the patient returning a completed questionnaire (two sample \(t\)

**Main findings**

Table 3 shows the variance components for the six outcome
measures estimated with the hierarchical models. In all cases,
most of the variance in patient level scores was due to
differences in ratings of the same doctor by different patients.
For both of the doctor specific measures that we investigated
(doctors’ communication and trust and confidence in the doctor),
the variance due to differences between doctors was greater
than that attributable to differences between practices, whereas
the reverse was true for the other four, non-doctors specific,
measures. For each outcome measure, table 4 shows the
number of patients’ ratings needed to achieve the 0.7 and 0.8
reliability thresholds, judged by authorities to represent
minimum acceptable thresholds in postgraduate assessment
settings.27 A substantial majority of doctors received sufficient
scores to achieve reliable estimates of performance in
communication—all but two of the 105 general practitioners in
our sample received at least 27 patients’ communication scores,
and all but 10 received 46 or more (overall mean 71 scores per
doctor).

Figure 1 shows the estimated mean communication scores
for individual doctors and for practices as a whole. It illustrates
the extent to which the variation in mean communication scores
between individual doctors (within practices) was greater than
the variation between practices and suggests that within practice
variability in doctors’ scores was greater in the lower scoring
practices. Further analysis confirmed this: the within practice
standard deviation of general practitioners’ mean communication
scores was negatively correlated with the practice’s mean
communication score (Pearson’s \(r=-0.505\); \(P=0.010\)). Figure
2 shows the adjusted doctor level and practice level mean
scores for “cleanliness of the practice buildings” and highlights,
in contrast to figure 1, the minimal within practice variability
between general practitioners for this non-doctor-specific
measure.

**Discussion**

Our results show that measurement of patients’ experience at
the practice level can mask considerable variation between
doctors within a practice. Our findings suggest that higher
performing practices usually comprise higher performing
doctors. However, lower performing practices may contain doctors with communication scores ranging from poor to very good. When the focus of patients’ ratings is on non-doctor-specific practice related attributes (such as the cleanliness of the facilities), these are measured well at the practice level.

**Strengths and limitations of study**

This was a large study, with survey responses from 7721 patients relating to 105 doctors in 25 practices. Our stratified sampling strategy secured participation from doctors delivering care in practices with a range of summary scores for interpersonal skills after adjustment for case mix, and this improves generalisability to wider primary care contexts. Our use of a postal survey resulted in an average delay of just over two weeks between the patient’s consultation and their receipt of the questionnaire—substantially less than the six month reference timeframe adopted in the English GP Patient Survey. Furthermore, we believe a two week delay is unlikely to be a significant source of recall bias and, in any event, would reasonably be expected to affect all doctors equally in this study.

Because of the quota sampling strategy, a simple participation rate for practices could not be calculated. Low response rates are commonly encountered in patient feedback surveys. However, our patient response rate of 51% was considerably higher than the 35% achieved in the most recent published results for the GP Patient Survey (July 2012 to March 2013) and in similar surveys elsewhere. Surveys such as these are typically used to inform national metrics of healthcare performance. Substantial variation existed in survey response rates between the general practitioners in our study, even after we accounted for the role of chance. Non-response will tend to inflate doctors’ and practices’ scores, but this inflation will be largest when non-response is highest. Given that the lowest response rates tend to occur for practices with the lower scores, any non-response bias would tend to attenuate the extent of variance between both doctors and practices rather than inflate it. The estimated magnitudes of such effects are small, and we would not expect them to alter the ratios of variance at the practice and doctor levels or affect our conclusions regarding the comparison of doctor and practice level variances.

Our sampling of practices from the lowest, middle, and highest quarter of GP Patient Survey practice level communication scores may mean that our estimate of the total amount of practice level variance could differ slightly from that of the full population; however, we believe that this does not affect our conclusions regarding the relation between practice level scores and the extent of within practice variation. Finally, we were blinded to patients’ postcodes and hence could not adjust our outcome measures for neighbourhood level deprivation. This limitation is unlikely to have biased our results, as we have previously shown that after sex, age, ethnicity, and health status are controlled for (as we did in this study), deprivation has a very small association with patients’ experience.

**Research in context**

Whereas Howie and colleagues described the variation within a sample of Scottish practices in respect of doctors’ communication, several other studies have used hierarchical models to apportion the sources of variance in patients’ feedback in primary care settings. Few studies have attempted to distinguish the relative contributions of doctors and organisations. In general, these studies concur with our finding that the proportion of variance due to doctors is greater than that due to practices in the case of doctor specific measures and is less in the case of non-doctor specific measures. In contrast to these findings, Rodriguez and colleagues found a greater proportion of variance in all types of measure, including physician-patient communication, to be due to sites, medical groups, and primary care service areas (combined) than due to doctors. Whether a distinction was made between doctors and organisations, all of these studies showed that most of the variance in patient level scores can be attributed to patients and residual sources. Salisbury and colleagues noted that a high proportion of the variance in communication scores in English general practice is attributable to patients and other factors, rather than to practices or individual doctors. The authors interpreted this as indicating that “so little variation exists at the level of the doctor that the reliability of using this type of measure to assess an individual doctor’s performance is questionable.” This interpretation ignores the fact that doctors (and practices) are assessed not by using the rating provided by a single patient but by using the average of many patients’ ratings. This considerably reduces the “noise” created by variation at the patient level. We suggest that the focus for survey data should be on unit level reliability: the proportion of variance in reporting unit samples’ means (for example, practices’ means or doctors’ means) attributable to true variation between units. Our results suggest that despite the high proportion of patient level variance in communication scores, for this survey instrument a reliable (>0.8) adjusted mean score for individual doctors can be obtained with 46 patient scores per general practitioner, so that only a small minority of variance in reported doctor level scores is attributable to patients and residual sources. This is in line with our previously published data examining patients’ feedback for the purposes of revalidation. With sample sizes smaller than this, a trade off must be made between reliability and the utility of conducting individual rather than group level evaluations.

The trade-off between the assessment and reporting of performance indicators at the level of either the organisation or the individual practitioner may be informed by considering both the nature of the indicator and sources of variance. We have shown that, for indicators that are most likely to be under the control of individual practitioners (such as doctor-patient communication), more variance is explained by doctors than by practices. This can be taken as a validation of the use of these indicators to measure the performance of individual doctors. Conversely, some indicators (such as the cleanliness of a practice) were observed to have more variance at the practice level. For such indicators, organisations are in control, and these indicators are more suitable for the evaluation of performance at the level of the organisation. Our findings suggest that current practice level performance indicators, although they provide a potentially useful overview of average performance, may not provide meaningful information to commissioning, providers, or users for some key domains, such as communication skills. In particular, practices singled out as having lower performance through assessments of doctor-patient communication aggregated at a practice rather than an individual level are likely to contain a range of doctors. Patients attending such practices may see a general practitioner with excellent interpersonal skills or, alternatively, may see a doctor who is less proficient at communicating with patients. Conversely, the assessment of communication at the practice level may mask quite how poorly some general practitioners perform, as excellent doctors will pull up the average practice score. This has important implications for the ability to manage the performance of practices, as inadequate interpersonal skills might be missed.
Finally, it is worth observing that patients may express choice through requests for continuity of care with a preferred doctor. In compiling performance indicators to inform patients’ choices of providers, it would be preferable to report communication scores at the individual practitioner level or to report the range of individual practitioners’ scores within an organisation where that can be done reliably. Communication is a key driver of overall patients’ satisfaction,11 and ensuring patients’ ability to access accurate information on performance is important if they are expected to make informed choices among providers. If quality indicators are to be used to identify poor performance rather than to inform patients’ choice, an alternative to the potentially costly option of obtaining communication scores for all individual practitioners might be to use organisation level assessments such as those provided in the English GP Patient Survey to screen for lower performing practices. Where potential concerns about performance are identified by this mechanism, individual level assessments could then be targeted to those organisations alone. This approach, considering a low practice level score as a high sensitivity but low specificity test of whether a particular doctor in the practice may have a lower score in respect of communication, may be worthy of consideration, although determining reliable threshold scores for such a test would need data from a much larger sample than was available in this study. In addition, the cost effectiveness of such an approach remains to be determined, having never been explored in detail. Further research would be useful to explore the feasibility and practicality of alternative approaches to generating performance data on doctor-patient communication. More widely, many unanswered questions remain about the association between the publication of performance data and quality improvement, including the mechanisms underpinning any personal or organisational changes precipitated, and the perspectives of users, providers, and commissioners about the expected utility of alternative approaches.

Conclusions

Current approaches to evaluating performance in communication frequently assess publicly reported indicators at an aggregate level, rather than enabling patients and other stakeholders to evaluate individual practitioners directly. Reporting communication related performance indicators at practice level may mask large variation between individual practitioners. Practice level surveys may offer potential to act as an initial screen for concerns about performance, with subsequent data gathering focusing on individual doctor level surveys in lower performing practices.

We thank the patients, practice managers, general practitioners, and other staff of the general practices who kindly agreed to participate in this study and without whom the study would not have been possible. Thanks also go to Emily Taylor, Jenny Newbould, Emma Whitton, Amy Grattan, Charlotte Paddison, and Dawn Swannott for invaluable help with study set-up, practice recruitment, data collection, and data entry. We also thank the Improve Advisory Group for their input and support throughout this study. MJR is now senior psychiatrist at the University of Plymouth Peninsula Schools of Medicine and Dentistry, Plymouth, UK.

Contributors: MJR wrote the statistical analysis plan, monitored data collection in the south west of England, analysed the data, and drafted and revised the paper. JLC designed the study, wrote the statistical analysis plan, oversaw the conduct of the study, and drafted and revised the paper. GAA contributed to the statistical analysis plan and revised the paper. MOR designed the study, oversaw the conduct of the study, and revised the paper. JLC carried out and monitored data collection in the east of England/London, oversaw the conduct of the study, and drafted and revised the paper, JLC is the guarantor.

Funding: This work was funded by a National Institute for Health Research Programme Grant for Applied Research (NIHR PGIAR) programme (RP-PG-0608-10050). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/col_disclosure.pdf (available on request from the corresponding author) and declare: the study was funded by the UK NIHR as an unrestricted research award; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Approval for the study was obtained from the South West 2 Research Ethics Committee on 28 January 2011 (ref: 09/H0202/65). Return of a completed questionnaire was taken to indicate patients’ consent to participate in the study.

Declaration of transparency: The lead author (study guarantor) affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data sharing: No additional data available.


5 Iacobucci G. General practices could be subject to “Ofsted style” ratings within two years. BMJ 2013;346:f1941.


What is already known on this topic

Publication of performance data is increasingly the norm in healthcare systems, although evidence to support the mechanisms by which such publication may drive quality improvement is variable.

Data collected to derive performance indicators may be collected at organisational (hospital, general practice) or individual practitioner level, with organisational level the norm for English general practice.

For practitioner level performance indicators, such as the communication skills of individual doctors, variance within organisations is often greater than that between organisations.

What this study adds

Practice level ratings of general practitioners’ consultation communication skills can mask considerable variation between the doctors within a practice, particularly in poorer performing practices.

Patients registered with practices that have communication scores at the lower end of the spectrum may experience wide variation in communication skills between individual doctors.

Higher scoring practices are very unlikely to include lower scoring doctors, but lower scoring practices require further investigation at the individual doctor level to distinguish higher and lower scoring general practitioners.


Accepted: 24 September 2014

Cite this as: BMJ 2014;349:g6034

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/.
## Tables

### Table 1 | Practice profiles and questionnaire response rates

<table>
<thead>
<tr>
<th>Setting</th>
<th>Banding on 2009/10 GPPS communication score*</th>
<th>GP head count</th>
<th>Participating doctors</th>
<th>List size (000s)</th>
<th>Deprivation index†</th>
<th>Overall response rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner city</td>
<td>Low</td>
<td>2</td>
<td>2</td>
<td>6.9</td>
<td>26.6</td>
<td>37.9</td>
</tr>
<tr>
<td>Inner city</td>
<td>Low</td>
<td>3</td>
<td>3</td>
<td>5.1</td>
<td>48.5</td>
<td>36.8</td>
</tr>
<tr>
<td>Inner city</td>
<td>Low</td>
<td>4</td>
<td>4</td>
<td>5.1</td>
<td>36.6</td>
<td>37.8</td>
</tr>
<tr>
<td>Inner city</td>
<td>Low</td>
<td>5</td>
<td>4</td>
<td>7.8</td>
<td>26.1</td>
<td>50.5</td>
</tr>
<tr>
<td>Inner city</td>
<td>Low</td>
<td>8</td>
<td>6</td>
<td>8.7</td>
<td>32.4</td>
<td>43.5</td>
</tr>
<tr>
<td>Inner city</td>
<td>Middle</td>
<td>2</td>
<td>2</td>
<td>2.5</td>
<td>30.1</td>
<td>47.0</td>
</tr>
<tr>
<td>Inner city</td>
<td>Middle</td>
<td>3</td>
<td>3</td>
<td>5.4</td>
<td>13.7</td>
<td>67.7</td>
</tr>
<tr>
<td>Inner city</td>
<td>Middle</td>
<td>6</td>
<td>6</td>
<td>8.0</td>
<td>39.4</td>
<td>32.0</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>2</td>
<td>2</td>
<td>3.5</td>
<td>15.2</td>
<td>71.0</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>2</td>
<td>2</td>
<td>2.9</td>
<td>22.2</td>
<td>58.9</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>2</td>
<td>2</td>
<td>3.2</td>
<td>29.6</td>
<td>24.1</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>3</td>
<td>3</td>
<td>6.6</td>
<td>15.1</td>
<td>55.8</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>4</td>
<td>4</td>
<td>4.1</td>
<td>18.3</td>
<td>59.3</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>5</td>
<td>5</td>
<td>12.0</td>
<td>27.6</td>
<td>58.9</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>5</td>
<td>5</td>
<td>6.0</td>
<td>19.3</td>
<td>52.6</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>7</td>
<td>6</td>
<td>9.7</td>
<td>20.0</td>
<td>53.8</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>8</td>
<td>7</td>
<td>16.5</td>
<td>14.4</td>
<td>45.1</td>
</tr>
<tr>
<td>Urban</td>
<td>Low</td>
<td>9</td>
<td>8</td>
<td>11.8</td>
<td>16.4</td>
<td>48.1</td>
</tr>
<tr>
<td>Urban</td>
<td>Middle</td>
<td>3</td>
<td>3</td>
<td>5.3</td>
<td>20.8</td>
<td>67.6</td>
</tr>
<tr>
<td>Urban</td>
<td>High</td>
<td>6</td>
<td>5</td>
<td>8.5</td>
<td>22.1</td>
<td>47.2</td>
</tr>
<tr>
<td>Urban</td>
<td>High</td>
<td>8</td>
<td>8</td>
<td>14.2</td>
<td>18.9</td>
<td>64.4</td>
</tr>
<tr>
<td>Rural</td>
<td>Middle</td>
<td>5</td>
<td>4</td>
<td>5.1</td>
<td>23.1</td>
<td>60.5</td>
</tr>
<tr>
<td>Rural</td>
<td>High</td>
<td>3</td>
<td>2</td>
<td>2.4</td>
<td>18.9</td>
<td>49.8</td>
</tr>
<tr>
<td>Rural</td>
<td>High</td>
<td>4</td>
<td>4</td>
<td>5.4</td>
<td>11.5</td>
<td>75.5</td>
</tr>
<tr>
<td>Rural</td>
<td>High</td>
<td>5</td>
<td>5</td>
<td>9.1</td>
<td>4.8</td>
<td>71.7</td>
</tr>
<tr>
<td>All</td>
<td>—</td>
<td>114</td>
<td>105</td>
<td>—</td>
<td>—</td>
<td>50.9</td>
</tr>
</tbody>
</table>

GPPS=General Practice Patient Survey.

*Low=below 25th centile; middle=between 37.5th and 62.5th centiles; high=above 75th centile.

†Average taken across practice population; these scores underlie figures reported by Public Health England at http://fingertips.phe.org.uk/profile/general-practice.
### Table 2: Demographic profile of responding patients (n=7721)

<table>
<thead>
<tr>
<th>Category</th>
<th>No (% of non-missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>4785 (62.4)</td>
</tr>
<tr>
<td>Male</td>
<td>2882 (37.6)</td>
</tr>
<tr>
<td>Missing</td>
<td>54</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;18</td>
<td>5 (0.1)</td>
</tr>
<tr>
<td>18-24</td>
<td>249 (3.2)</td>
</tr>
<tr>
<td>25-34</td>
<td>786 (10.3)</td>
</tr>
<tr>
<td>35-44</td>
<td>983 (12.8)</td>
</tr>
<tr>
<td>45-54</td>
<td>1150 (15)</td>
</tr>
<tr>
<td>55-64</td>
<td>1474 (19.2)</td>
</tr>
<tr>
<td>65-74</td>
<td>1550 (20.2)</td>
</tr>
<tr>
<td>75-84</td>
<td>1171 (15.3)</td>
</tr>
<tr>
<td>≥85</td>
<td>299 (3.9)</td>
</tr>
<tr>
<td>Missing</td>
<td>54</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>White British</td>
<td>6138 (81.5)</td>
</tr>
<tr>
<td>White Irish</td>
<td>132 (1.8)</td>
</tr>
<tr>
<td>Any other white background</td>
<td>459 (6.1)</td>
</tr>
<tr>
<td>Mixed white and black Caribbean</td>
<td>23 (0.3)</td>
</tr>
<tr>
<td>Mixed white and black African</td>
<td>10 (0.1)</td>
</tr>
<tr>
<td>Mixed white and Asian</td>
<td>18 (0.2)</td>
</tr>
<tr>
<td>Any other mixed background</td>
<td>19 (0.3)</td>
</tr>
<tr>
<td>Asian or Asian British—Indian</td>
<td>169 (2.2)</td>
</tr>
<tr>
<td>Asian or Asian British—Pakistani</td>
<td>55 (0.7)</td>
</tr>
<tr>
<td>Asian or Asian British—Bangladeshi</td>
<td>71 (0.9)</td>
</tr>
<tr>
<td>Any other Asian background</td>
<td>72 (1)</td>
</tr>
<tr>
<td>Black or black British—Caribbean</td>
<td>95 (1.3)</td>
</tr>
<tr>
<td>Black or black British—African</td>
<td>161 (2.1)</td>
</tr>
<tr>
<td>Any other black background</td>
<td>9 (0.1)</td>
</tr>
<tr>
<td>Chinese</td>
<td>45 (0.6)</td>
</tr>
<tr>
<td>Any other ethnic group</td>
<td>57 (0.8)</td>
</tr>
<tr>
<td>Missing</td>
<td>188</td>
</tr>
<tr>
<td><strong>Health</strong></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>714 (9.5)</td>
</tr>
<tr>
<td>Fair</td>
<td>1827 (24.3)</td>
</tr>
<tr>
<td>Good</td>
<td>2502 (33.2)</td>
</tr>
<tr>
<td>Very good</td>
<td>1961 (26.1)</td>
</tr>
<tr>
<td>Excellent</td>
<td>523 (6.9)</td>
</tr>
<tr>
<td>Missing</td>
<td>194</td>
</tr>
</tbody>
</table>
Table 3 | Percentages of variance in adjusted mean outcome scores that are attributable to practices, doctors, and patients

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Source of variance</th>
<th>Practice</th>
<th>Doctor</th>
<th>Patients and residual error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication score</td>
<td></td>
<td>1.8</td>
<td>6.4</td>
<td>91.9</td>
</tr>
<tr>
<td>Confidence and trust</td>
<td></td>
<td>0.8</td>
<td>5.2</td>
<td>94.0</td>
</tr>
<tr>
<td>Overall satisfaction</td>
<td></td>
<td>6.0</td>
<td>1.1</td>
<td>92.9</td>
</tr>
<tr>
<td>Helpfulness of</td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>92.2</td>
</tr>
<tr>
<td>Cleanliness of health</td>
<td></td>
<td>10.6</td>
<td>0.3</td>
<td>89.1</td>
</tr>
<tr>
<td>Ease of getting into</td>
<td></td>
<td>1.9</td>
<td>0.4</td>
<td>97.6</td>
</tr>
</tbody>
</table>
Table 4  Number of patients’ ratings needed to achieve reliability of 0.7 or 0.8 for doctor’s raw and adjusted mean scores

<table>
<thead>
<tr>
<th>Communication score</th>
<th>Confidence and trust</th>
<th>Overall satisfaction with surgery</th>
<th>Helpfulness of receptionists</th>
<th>Cleanliness of health centre</th>
<th>Ease of getting into building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability of raw mean score</td>
<td>0.7</td>
<td>21</td>
<td>30</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>36</td>
<td>51</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>Reliability of adjusted mean score*</td>
<td>0.7</td>
<td>27</td>
<td>37</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>46</td>
<td>63</td>
<td>53</td>
<td>48</td>
</tr>
</tbody>
</table>

*Adjusted for patient’s sex, age, ethnicity, and self reported health status.
Figures

**Fig 1** Mean communication score (best estimate) by practice and doctor. Practices (n=25) are sorted by their mean communication score. Horizontal shading serves only as visual separation of results for different practices. Reliability calculations using variance components showed that achieving acceptable reliability (>0.7) for general practitioners’ adjusted mean communication scores with 27 patients’ scores and good reliability (>0.8) with 46 patients’ scores per doctor is feasible (see appendix). All but 10 of the 105 participating doctors had more than 46 scores; two received less than 27 scores (mean 71 scores per doctor). Data for these doctors was retained in the subsequent modelling, as use of best linear unbiased predictors to estimate doctors’ mean scores has a “conservative” effect. Where sample sizes are smaller, estimated mean scores are drawn closer to practice mean.

**Fig 2** Mean score for cleanliness of practice building (best estimate) by practice and doctor. Practices (n=25) are sorted by their mean score for cleanliness. Horizontal shading serves only as visual separation of results for different practices.