BMJ

BMJ 2014;348:g330 doi: 10.1136/bmj.g330 (Published 27 January 2014)

RESEARCH

Withdrawing performance indicators: retrospective analysis of general practice performance under UK Quality and Outcomes Framework


OPEN ACCESS

Evangelos Kontopantelis *senior research fellow*[1 2], David Springate *research associate*[1 3], David Reeves *reader*[1 3], Darren M Ashcroft *professor*[4], Jose M Valderas *professor*[5], Tim Doran *professor*[6]

[1]NIHR School for Primary Care Research, Centre for Primary Care, Institute of Population Health, University of Manchester, Manchester M13 9PL, UK; [2]Centre for Health Informatics, Institute of Population Health, University of Manchester; [3]Centre for Biostatistics, Institute of Population Health, University of Manchester; [4]Centre for Pharmacoepidemiology and Drug Safety Research, Manchester Pharmacy School, University of Manchester; [5]Institute for Health Services Research, UE Medical School, University of Exeter, Exeter, UK; [6]Department of Health Sciences, University of York, York, UK

Abstract

Objectives To investigate the effect of withdrawing incentives on recorded quality of care, in the context of the UK Quality and Outcomes Framework pay for performance scheme.

Design Retrospective longitudinal study.

Setting Data for 644 general practices, from 2004/05 to 2011/12, extracted from the Clinical Practice Research Datalink.

Participants All patients registered with any of the practices over the study period—13 772 992 in total.

Intervention Removal of financial incentives for aspects of care for patients with asthma, coronary heart disease, diabetes, stroke, and psychosis.

Main outcome measures Performance on eight clinical quality indicators withdrawn from a national incentive scheme: influenza immunisation (asthma) and lithium treatment monitoring (psychosis), removed in April 2006; blood pressure monitoring (coronary heart disease, diabetes, stroke), cholesterol concentration monitoring (coronary heart disease, diabetes), and blood glucose monitoring (diabetes), removed in April 2011. Multilevel mixed effects multiple linear regression models were used to quantify the effect of incentive withdrawal.

Results Mean levels of performance were generally stable after the removal of the incentives, in both the short and long term. For the two indicators removed in April 2006, levels in 2011/12 were very close to 2005/06 levels, although a small but statistically significant drop was estimated for influenza immunisation. For five of the six indicators withdrawn from April 2011, no significant effect on performance was seen following removal and differences between predicted and observed scores were small. Performance on related outcome indicators retained in the scheme (such as blood pressure control) was generally unaffected.

Conclusions Following the removal of incentives, levels of performance across a range of clinical activities generally remained stable. This indicates that health benefits from incentive schemes can potentially be increased by periodically replacing existing indicators with new indicators relating to alternative aspects of care. However, all aspects of care investigated remained indirectly or partly incentivised in other indicators, and further work is needed to assess the generalisability of the findings when incentives are fully withdrawn.

Introduction

As part of wider efforts to improve the quality and efficiency of healthcare, purchasers worldwide have experimented with linking performance indicators to financial incentives, reputational incentives, or both, within pay for performance and public reporting schemes. As the clinical evidence base and policy priorities change over time, indicator sets must be periodically reviewed and individual indicators modified, removed, or replaced. Within financial incentive schemes, indicators may also be removed because achievement rates have reached a ceiling, thereby allowing new indicators, for which improvement is possible, to be introduced.[1]

Incentives are intended to improve performance by changing physicians' behaviour, but even when this approach is successful the change may be temporary. If the incentive is necessary to maintain high performance levels, its withdrawal will result in lower achievement rates and a loss of performance gains. This may occur, for example, because better performance requires additional staffing resource that depends on the incentive payments or because physicians' expectations of reward are altered. Depending on the nature of the incentives and the extent

Correspondence to: E Kontopantelis e.kontopantelis@manchester.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/348/bmj.g330?tab=related#webextra)

of their negative effects on other motivations for providers, particularly intrinsic motivations, achievement may even fall below performance levels attained before incentivisation.[2] Alternatively, if incentives increase the perceived priority of the activities,[3] support the establishment of quality improvement infrastructures and practices, or habituate providers to perform at a high level, then achievement rates might be maintained after withdrawal of the incentive. Such normalisation would require the relevant processes and behaviours to become so routinely embedded and integrated into providers' practice that the incentives become superfluous.[4]

To date, few examples of indicators being withdrawn from incentive schemes have been seen, so evidence on the effects is limited. When incentives to screen patients for diabetic retinopathy and cervical cancer were withdrawn from a Kaiser Permanente scheme in California, achievement rates fell by 3.1% and 1.6% a year respectively.[5] These losses exceeded the gains made during the preceding incentivisation period.

In the United Kingdom, the Quality and Outcomes Framework (QOF) incentive scheme provides family practices with financial rewards linked to performance on a range of more than 100 quality of care indicators, mostly related to processes of care for common chronic conditions.[6 7] Practices can exclude ("exception report") patients deemed inappropriate from the payment calculations for various reasons (for example, intolerance to a specified treatment or informed dissent by the patient).[8] The overall annual cost of the QOF exceeds £1bn per annum,[9] and the scheme has increased the average annual income of non-salaried general practitioners by £23 000 (€27 640; $37 580) (approximately 30% of the average pre-incentivisation income of £75 000).[7] Performance on quality indicators is recorded on practices' clinical computing systems and is centrally monitored through the national Quality and Management Analysis System database.

The QOF is reviewed annually in a process overseen by the National Institute for Health and Care Excellence, which makes recommendations for individual indicators to be modified or removed. Final agreement on changes to indicators is reached in negotiations between the Department of Health and the British Medical Association. For the third year of the QOF (2006/07), three clinical indicators were removed: one after the emergence of new evidence on the efficacy of treatment (influenza immunisation for patients with asthma)[10 11] and two that partially overlapped with other broader indicators (spirometry for new cases of chronic obstructive pulmonary disease and monitoring of lithium concentrations for patients on lithium treatment). For the eighth year (2011/12), a further eight indicators were removed specifically because achievement rates were judged to have reached a ceiling, even though the activities were still deemed to represent best practice.[1] The central Quality and Management Analysis System database does not monitor performance for removed indicators.

The aim of this study was to assess the effect that removing the incentives for these indicators from the QOF scheme had on subsequent performance, both on performance as measured by the same indicator and on performance as measured by related indicators.

Methods

Data source

Indicators removed from the QOF scheme are not routinely measured and reported after removal. To investigate the effect of the withdrawal of the incentive on practices' performance, we reconstructed the relevant indicators by using a large primary care database, the Clinical Practice Research Datalink (CPRD). This database holds complete electronic patients' records (including diagnoses, prescriptions, and referrals) from participating general practices with the Vision clinical computer system, used in approximately a fifth of all English practices.[12] Patients' data are recorded in the form of Read codes—a hierarchical clinical coding system. In July 2012 data were available for 644 practices and 13 772 992 patients. Figure 1⇓ shows details on complete patients' records within and across the study period.

Characteristics of final datasets

The study period extended from 1 April 2004, the date of introduction of the incentivisation scheme, to 31 March 2012. Practices' performance under the QOF is measured over a financial year, so we divided the study period into eight financial years (1 April to 31 March the following year). Not all 644 practices provided research standard data (as assessed by the CPRD assessment algorithm) for the whole period. Within each year, we identified practices that reliably contributed data for the whole year. Our main dataset comprised this group of practices, which varies over time. We also generated two alternative datasets with which to assess the sensitivity of our findings. For the first, we included 452 practices that were continuously active and up to standard for the whole of the study period; for the second, we selected a subsample of 50 practices that were most representative of UK practices in terms of list sizes of patients and area deprivation according to the Index of Multiple Deprivation,[13 14] two of the most important predictors of QOF performance.[12 15 16] In each of the three datasets, for each financial year, we defined "eligible" patients as those registered with an included practice for the full year. Figure 1⇓ describes the process in detail, and table 1⇓ shows the available characteristics of the practices (patients and practices are anonymised in the CPRD).

Conditions

We chose seven chronic conditions for which quality indicators had been included in, and subsequently removed from, the QOF scheme—asthma, coronary heart disease, chronic obstructive pulmonary disease, diabetes mellitus, epilepsy, severe mental health (psychotic illness), and stroke—and one for which a quality indicator measuring a process similar to the process incentivised in a removed indicator was available—hypertension. To identify patients with each condition in the CPRD, we used the QOF business rule code sets (the algorithms used for the identification of patients in this incentive scheme) in addition to relevant keywords identified by clinicians to generate unrefined, inclusive lists of Read codes and other clinical activity codes. This more inclusive approach aimed to account for changes in the business rules over time and the dynamic nature of code usage. Two clinicians independently reviewed these lists and reached consensus on a conservative list of codes (indicating the presence of the respective condition with a high degree of certainty). For diabetes, for example, Read code C107.12 (diabetes with gangrene) was included and 13B1.00 (diabetic diet) was excluded. Read codes used in the study are available from the clinical codes repository.[17] We treated all conditions, except asthma, as chronic and unresolvable, so that we considered a patient with a relevant code at any point during the study period to have the condition from that time onwards. Patients with asthma with a code denoting resolution of the condition were excluded from the denominator (the set of patients designated

to have the condition) from the date of the resolution code. To comply with QOF definitions, we limited denominators for diabetes and epilepsy to patients aged 17 or over and 18 or over respectively.

Characteristics of removed indicators

In the first eight years of the incentivisation scheme, 11 indicators were removed (see table 2⇓ and appendix table A1). Ten of these indicators had been introduced in year 1 of the scheme (2004/05), and one (MH7) had been introduced in year 3 (2006/07). Ten of the removed indicators related to the monitoring of particular aspects of patients' care; one indicator related to the treatment provided (influenza immunisation for adult patients with asthma). Seven of the monitoring indicators related to physiological or biochemical measurements (such as a record of blood pressure), and for each of these we also modelled the corresponding intermediate outcome indicator from the QOF scheme (for example, blood pressure ≤150/90 mm Hg).

We modelled the indicators in the CPRD mainly by using Read codes, but we also included codes relating to drugs, tests, and test results where appropriate. For example, we used codes for administered influenza immunisation products in addition to appropriate Read codes to model the influenza immunisation indicators, and we used test values for the intermediate outcome indicators. We also modelled several additional, unremoved, indicators to use as covariates in the analyses, which are also shown in table 2⇓ (see statistical modelling section).

Although some indicators have undergone small or moderate changes since their introduction, we used a single static definition to reconstruct each in the CPRD, to more reliably model changes in performance over time (appendix table A1). To construct each indicator, we defined relevant numerators and denominators. For example, for indicator Asthma7 (percentage of patients aged 16 and over with asthma who had influenza immunisation in preceding 1 September to 31 March) we defined the denominator as the number of patients with asthma in the relevant financial year and the numerator as the number of those patients who were immunised between 1 September and 31 March of the same financial year. For intermediate outcome indicators, we limited the denominators to patients for whom we were able to extract at least one non-missing test value in the defined period (usually 15 months) and the numerator to the subgroup of patients whose last recorded test value was within the range required by the indicator.

We report on indicators that we successfully constructed, on the criterion of exhibiting scores and trends comparable to those reported under the QOF. Our a priori decision was to discard indicators that could not be modelled reliably. However, comparison of the scores on our constructed indicators with those reported under the QOF (through the Quality and Management Analysis System) could only be approximate. Under the QOF, practices are allowed to "exception report" (exclude) patients from care, and hence from calculation of scores on the indicators, for a variety of clinical or logistical reasons.[8] We included these patients in the modelled indicators, to avoid potential bias should exception reporting rates themselves change as a result of removal of indicators,[18] focusing on a population measure of quality that is free from potential manipulation.

Statistical modelling

We did two sets of analyses, using multilevel multiple linear regressions and a longitudinal interrupted time series design. The first set of analyses examined whether the removal of an indicator from the incentives framework affected the subsequent mean performance of practices as measured by that indicator. The second set of analyses investigated the effect of the removal of each monitoring indicator on the corresponding intermediate outcome indicator.

On examination, the levels and trends of the indicators related to medication review in patients with epilepsy (EPI3/7), follow-up of severe mental health disorders (MH7), and spirometry in new chronic obstructive pulmonary disease patients (COPD2) were assessed as unreliable and were not included in the analysis. For example, rates of spirometry were close to zero (compared with mean national reported rates under the QOF of more than 90%, for more than 750 000 patients), indicating that the relevant Read codes were systematically not captured in the version of the CPRD that we used. The levels and trends of the indicators accepted as reliable were comparable to levels reported nationally under the QOF,[19] although levels were lower because of the inclusion of exception reported patients.

For withdrawn indicators, to quantify the effect of removal by 2011/12, we used multilevel regression models to generate practice level predictions based on the pre-intervention level and trend of the withdrawn indicator in the previous three years (two if removed in April 2006). To better account for the variation in performance levels over time and the changes in our sample, we controlled the predictions for performance on identical process indicators in other disease groups (if available), performance on similar process indicators within the same disease group, and practices' characteristics. We then subtracted the post-removal model estimates from the observed scores and used a meta-analysis method to combine them across practices into an overall "removal" effect.[20] Table 2⇓ describes the design and the indicators used. For "linked" outcome indicators, the approach was the same but we did not control for other outcome indicators within the disease group because we did not identify any that we considered similar.

Before implementation, we validated the method for short term and long term effects of removal. For the short term predictions (that is, 2011/12 when the indicator was removed in April 2011), we assumed that indicators were removed in April 2010 and used the method to predict 2010/11, hypothesising that the overall effect would be very close to zero across all models. We found that to be the case, and, although small changes in the specification of the models did not affect the results greatly, the inclusion of the control indicators improved overall performance. For the long term predictions (that is, 2011/12 when the indicator was removed in April 2006), we used indicators that were not withdrawn before 2010/11 but assumed that they were withdrawn in April 2006 to estimate the performance of the models in 2010/11, again hypothesising that we would not observe removal effects. However, we did observe moderate effects in some models, and the obtained results were very sensitive to small changes in the specifications. Therefore, we decided not to use this predictions-observations comparison method for the long term investigation; instead, we made a simple comparison between performance levels in the last time point pre-removal (2005/06) and the levels in 2011/12, controlling for practices' characteristics in a multilevel regression analysis. The full details of the modelling are provided in the web appendix.

For all main analyses, we logit transformed indicator scores to account for potential ceiling effects and the variation in effort needed to increase performance at different levels; that is, we assumed that, for example, more effort is required to affect an improvement from 90% to 95% than for an improvement from 60% to 65%. This non-linear relation is modelled through the transformed score.[21] The analysis on the transformed scores also ensures that predictions fall within the 0-100 range. In instances where a practice score was at 100% or 0% (resulting in a transformed score of $+\infty$ or $-\infty$ respectively), we applied the empirical logit.[22] For better interpretability, we present predicted scores and differences (from observed) that are back transformed to percentages. For indicators on which some practices scored either 0% or 100%, the back transformed practice mean does not correspond exactly to the mean calculated using untransformed data. We used Stata v12.1 for all analyses.

We repeated all analyses on two subsamples of the main dataset (fig 1⇓) and using untransformed indicator scores. We present results for three of the five sensitivity analyses (sensitivity dataset 1 and logit scores; sensitivity dataset 2 and logit scores; main dataset and untransformed scores) in the appendix and discuss differences in the results section.

Results

The practices included in the study were broadly representative of English practices with respect to area deprivation but tended to be much larger on average than practices nationally. In addition, practices from the North East, Yorkshire and the Humber, and East Midlands regions were under-represented in the database (table 1⇓).

Disease prevalence rates calculated using the database were broadly comparable to rates reported under the QOF (table 3⇓). Recorded prevalence rates declined for asthma and coronary heart disease and increased for chronic obstructive pulmonary disease and diabetes over the study period. Recorded prevalence rates for hypertension, epilepsy, psychosis, and stroke remained relatively stable. Levels and trends were largely unchanged when calculated on the two sensitivity samples (appendix table A2)

Performance on indicators

Indicators removed in April 2006

For Asthma7 (patients with asthma receiving influenza immunisation), mean performance remained relatively stable across the incentivisation (2004/05 to 2005/06) and post-incentivisation (2006/07 to 2011/12) periods, ranging from 78.0% to 79.0%. In comparison, mean performance on the four influenza immunisation indicators that remained in the scheme was higher throughout the entire study period, remaining stable between 2004/05 and 2007/08, before deteriorating somewhat in later years.

For MH3 (patients on lithium treatment with a record of lithium concentrations), mean performance improved from 91.1% in 2005/06 (the last year the indicator was included in the scheme) to 92.5% in 2011/12. Performance on the corresponding intermediate outcome indicator (MH5/18: patients on lithium treatment with lithium concentrations in the therapeutic range) was also quite stable from 2005/06 onwards.

Indicators removed in April 2011

For the blood pressure monitoring indicators removed in April 2011 (CHD5, DM11, and Stroke5), average performance remained high after removal and very close to levels in previous years (92-94%). Performance for the blood pressure monitoring indicator that remained in the scheme (BP4: monitoring in hypertensive patients) also remained stable at around 90%. Performance on each of the corresponding intermediate outcome indicators (control of blood pressure) improved throughout the study period.

For the cholesterol monitoring indicators, a small decline in mean performance was apparent for CHD7 (from 88.3% in 2010/11 down to 87.0% in 2011/12) but DM16 showed stability (91.4% in 2010/11 and 91.2% in 2011/12). Performance for Stroke7, the only cholesterol monitoring indicator that remained in the scheme, also seemed stable, at 85.3% in 2010/11 and 85.5% in 2011/12. Performance for the cholesterol intermediate outcome indicators CHD8 and Stroke8 seems to have dropped very slightly in 2011/12 compared with the previous few years, whereas for DM17 the decrease was more pronounced, although mean performance had been slowly declining for several years.

Mean practice performance in monitoring HbA_{1c} measurements (DM5) remained stable at around 92% following the indicator's removal in 2011/12. Performance on the corresponding intermediate outcome indicator (DM6/20/23/26) increased until 2010/11 (71.4%) then fell back to 70.4% in 2011/12.

Effect of indicator removal

Tables 4⇓ and 5⇓ show findings from the short term comparison of observed performance after removal of an indicator with our estimates of the performance expected had the indicator not been removed. Results from the long term analyses are discussed below and provided in appendix table A8. Indicator scores and short term predictions are also plotted in figures 2⇓ and 3⇓. The values presented in table 5⇓ are results from the analysis of logit transformed indicator scores, back transformed into percentages. As such, practice means in table 5⇓ do not always correspond to the raw means given in table 4⇓.

Indicators removed in April 2006 and linked indicators

For Asthma7, the adjusted (controlled for practices' characteristics) back transformed mean difference between 2005/06 and 2011/12 levels was −0.70% (95% confidence interval −1.01% to −0.39%), indicating a very small drop in performance over time. The difference between 2005/06 and 2011/12 levels for MH3 was not statistically significant (0.65%, −0.11% to 1.46%). The linked intermediate outcome indicator MH5/18 (lithium concentrations within the therapeutic range) also showed no significant difference between 2005/06 and 2011/12 levels (0.63%, −0.38% to 1.72%), following removal of MH3.

Indicators removed in April 2011 and linked indicators

The indicators for monitoring blood pressure (CHD5, DM11, and Stroke5), HbA_{1c} (DM5), and cholesterol in patients with diabetes (DM16) all showed no statistically significance differences between observed and expected levels following removal. However, the cholesterol monitoring indicator for patients with coronary heart disease (CHD7) showed a significantly lower observed mean in 2011/12 compared with expectation (−1.19%, −1.56% to −0.81%).

For the linked indicators relating to blood pressure control, observed performance for CHD6 in 2011/12 was very close to expectation, and for DM12/30 and Stroke6 differences of around 0.3% were found, with only the last one reaching statistical significance (−0.35%, −0.65% to −0.05%). The two cholesterol

control indicators had observed mean scores in 2011/12 only slightly, but significantly, below expectation (CHD8: −0.32%, −0.62% to −0.02%; DM17: −0.45%, −0.75% to −0.15%), but we found a larger difference for the HbA_{1c} control indicator DM6/20/23/26 (−2.08%, −2.45% to −1.71%).

Sensitivity analyses

For the two indicators removed in April 2006 (Asthma7 and MH3), performance rates over time remained at least as high as in the pre-removal years. We observed a similar pattern for the indicators removed in April 2011. Levels of indicator scores were almost identical in sensitivity analysis 1 (all contributing practices across the whole time period) but slightly higher in sensitivity analysis 2 (50 more representative practices in terms of list size). Trends are given in appendix tables A3 and A4.

Results were broadly similar in all sensitivity analyses (appendix tables A5-A8). Estimates from sensitivity analysis 1 (sensitivity dataset 1 and logit scores) were similar to the ones obtained in the main analysis, and no differences existed in the conclusions. In sensitivity analysis 2 (sensitivity dataset 2 and logit scores), we found fewer statistically significant differences (for example, no differences for Asthma7, CHD7, CHD8, DM17, or Stroke6), reflecting the much smaller sample of practices. In sensitivity analysis 3 (main dataset but with untransformed scores), we found no statistically significant difference for Stroke6 (although the effect was of similar magnitude), but we observed statistically significant differences for DM6 (in 2011/12) and MH3.

Discussion

The recent proliferation of pay for performance schemes in healthcare reflects a perception in some policy circles of providers' motivation as self interested, and physicians and other professionals are increasingly induced with explicit incentives linked to quality metrics.[23] If physicians' behaviour is primarily self interested, financial and reputational incentives should be effective in improving performance, but only while the incentives are in place. Evidence from outside the health field suggests that extrinsic motivators such as financial incentives not only are transitory in their effects but can actually be damaging in the longer term: they can diminish intrinsic motivators, including professional and moral motivations, which may not recover once the extrinsic motivator is withdrawn.[24 25] Financial incentives can therefore be both expensive and, in the longer term, counterproductive.

For incentives in healthcare to buck this trend, the professional and altruistic motivations of providers would need to be more robust than those in other fields, or the incentives would have to be so carefully aligned that intrinsic motivations are reinforced (or at least, not damaged).[3] Alternatively, changes to infrastructure made by providers to attain quality targets—or resulting from reinvestment of rewards—could lead to sustained improvements in performance beyond the period of incentivisation. In this study, we modelled the effect of withdrawing a range of incentives on subsequent performance under a comprehensive, national scheme for primary care providers. For five of the six indicators withdrawn in 2011/12, we found no significant effect on subsequent short term performance. For one of the two indicators removed in April 2006, adjusted levels in 2011/12 were not significantly different from 2005/06 levels. However, estimated differences were relatively small across all indicators, including for the two indicators that showed statistically significant deterioration.

Strengths and limitations of study

The main strength of the study was its the use of millions of electronic medical records from hundreds of general practices (using the same information clinicians used for providing care for the patients, thereby minimising observer effects) to construct relevant quality indicators and evaluate the effect of withdrawal of incentives. However, some important limitations exist. Firstly, the withdrawn monitoring indicators we modelled remained incentivised through their linked outcome indicators that remained in the scheme, as "not measuring something in the required time" is counted as "failed to achieve relevant intermediate outcome target." A strong indirect incentive for taking these measures thus still exists. For this reason, greater effects on performance may be apparent for withdrawn measurement indicators without a linked incentivised outcome.

Secondly, indirect incentivisation of withdrawn indicators exists for certain subpopulations of patients (for example, for 2011/12, 18.8% of asthma patients aged 16 or over had at least one of the four comorbidities for which the influenza immunisation incentive was not withdrawn). We decided not to exclude these comorbid cases so that our modelled indicators would not differ in their populations from those defined under the QOF. In addition, UK practices are also incentivised through a different scheme to immunise patients aged 65 or over against influenza, further partially incentivising the asthma influenza indicator for approximately 25.2% of our patients in 2011/12. These figures for comorbidity and age broadly agree with what has been reported elsewhere.[26] However, for 2011/12, 67.3% of the patients in the denominator of the indicator were not indirectly affected by any form of comorbidity or age related incentive.

Thirdly, CPRD practices are broadly representative in terms of local area deprivation, but they tend to be larger than the average English practice and use a single clinical computing system (Vision 3, used in 19% of the 8200 plus English practices). Choice of clinical system is a predictor of QOF performance,[12] so the generalisability of our findings might be limited. Fourthly, although CPRD prevalence rates and trends generally agree with nationally reported rates (table 3⇓), some small differences exist that might indicate with election bias or a problem representativeness.

Fifthly, indicators have characteristics (such as points values/remuneration and payment thresholds) that might affect performance. However, these have remained relatively stable over time and their effects could not be accounted for in the models owing to collinearity. Sixthly, we used an interrupted time series design to quantify the removal effects. This method is arguably the best possible approach in the absence of a control group,[27] but it is sensitive to assumptions and we decided not to use it for the indicators removed in April 2006 as we would have had to extrapolate many years into the future.

Seventhly, we did not model exceptions, and for some patients the care represented by an indicator will be inappropriate. Nevertheless, we argue that the potential for bias or manipulation is greater if excepted patients are not included in the analyses. Eighthly, we used fixed definitions of indicators, but within the QOF scheme some indicators changed over time (for example, the target for DM6 (HbA_{1c} control) varied from 7.0% to 7.5%). However, we prioritised consistency for the time series analysis. Finally, we originally aimed to model the effects of both year and each indicator as varying by practice (that is, random effects), but these models were very complex and did not converge in some cases. We therefore modelled only year as a random effect.

Findings

Performance seems to have deteriorated modestly for one of the two indicators withdrawn in year 3 of the scheme (2006/07). For Asthma7 (influenza immunisation, worth up to £1527 for the average practice), immunisation rates seemed to be stable post-incentivisation, although the model estimated a very small drop between 2005/06 and 2011/12 levels. Asthma7 was withdrawn in the light of new evidence on the appropriateness of immunising patients with asthma,[10 11] so any relative decline in immunisation rates may be attributable to practices responding to new evidence rather than—or in addition to—the withdrawal of financial incentives. By 2011/12, six years after the withdrawal of incentives, immunisation rates were 0.6% higher than in 2005/06, the final year of incentivisation, although that small increase might be attributable to the changing characteristics of CPRD practices (the adjusted difference showed a 0.7% drop which translates to approximately 2.6 patients per practice and 21 500 patients nationally). However, the effect of incentive withdrawal was clearly minor at the level of individual practices. This stability in immunisation rates is somewhat unexpected, given the uncertainty about the efficacy of immunisation in this group of patients. Our estimated rates of achievement for all influenza immunisation indicators were approximately 10 percentage points higher than what has been reported under the QOF and elsewhere.[28] This discrepancy can be explained by our more inclusive definition of the intervention: we used both Read codes and influenza immunisation products to define immunisation, which we felt was more realistic in capturing exposure and avoiding potential coding bias, whereas only Read codes are used under the QOF. Nevertheless, we used the stricter "Read code only" definition to assess the sensitivity of our findings for influenza immunisation, and although levels were lower they were again stable over time for all indicators. However, we also observed numerous patients who were excluded from these indicators but for whom care was met—a finding that warrants further investigation.

In the first two years of the QOF, two indicators incentivised the monitoring of lithium concentrations: MH3 (measurement of lithium concentrations, worth up to £382 for the average practice) and MH5 (lithium concentrations within the therapeutic range, worth up to £636). In year 3 (2006/07), when indicator MH3 was withdrawn, the maximum remuneration for MH5 was reduced by 60% and the upper threshold (the level of performance required to secure maximum remuneration) was increased from 70% to 90%. Practices therefore had to work harder for less reward: maximum remuneration fell from £1018 to £255 for the average practice, and these lower rewards were attainable only if lithium concentrations were maintained within the therapeutic range. We did not, however, observe any deterioration in monitoring rates following withdrawal of the incentive: after a steep increase between 2004/05 and 2005/06, rates continued to increase more slowly between 2005/06 and 2011/12. It could be argued that rates would have increased more quickly under incentives had the initial momentum been sustained, but the observed trend for MH3 was consistent with performance on other measurement indicators maintained within the QOF scheme. For the linked control indicator (MH5/18: lithium concentrations within the therapeutic range), performance continued to improve between 2005/06 and 2008/09 before falling off, but it remained above 2005/06 levels.

For all the indicators removed in April 2011, levels of performance were high (over 85%) in the first year of the scheme and remained high for the next six years, which ultimately led to their withdrawal. These indicators were also linked to intermediate outcomes indicators, so some financial incentive was retained in the post-incentivisation period. For example, after removal of the blood pressure monitoring indicator for patients with diabetes (DM11), practices still needed to measure blood pressure to achieve the target for blood pressure control (DM12). For five of the six removed indicators we analysed, we found no significant change in achievement rates following removal of the direct incentives, and the high levels of performance were maintained. In the case of CHD7 (cholesterol monitoring in patients with coronary heart disease) measurement rates fell to 1.2% below projected rates, equivalent to approximately 2.6 missed patients in the average practice and more than 21 500 patients nationally. Of the indicators withdrawn from April 2011, CHD7 was subject to the joint highest incentive (£890 for the average practice), had the lowest baseline achievement rate (85.2% in 2004/05), and increased the most under incentivisation (by 3.1% in the first seven years). Practices seem to have had a greater response both to the introduction and to the removal of incentives for this activity, and this warrants further investigation.

Of the five intermediate outcome indicators that were linked with activities withdrawn in 2011/12, four scored below projections although levels of performance remained high: blood pressure control for stroke (Stroke6), cholesterol control for diabetes (DM17) and coronary heart disease (CHD8), and glucose control for diabetes (DM6/20/23/26). However, differences between observed and expected performance were very small except for DM6/20/23/26, the indicator with the most changes in definition over time—a fact that probably partly explains the finding. The HbA_{1c} threshold for DM6/20/23/26 changed from 7.4% for 2004/05-2005/06 to 7.5% for 2006/07-2008/09, to 7% for 2009/10-2010/11, and back to 7.5% for 2011/12 (appendix table A1).

Implications and conclusions

The success of incentive schemes depends not only on their effect on providers' performance while incentives are active but also on subsequent performance once incentives are withdrawn. English practices achieved modest improvements in performance across a wide range of clinical activities under the substantial incentives in the Quality and Outcomes Framework. Following the withdrawal of incentives for several activities, levels of performance were generally maintained (including influenza immunisation for asthma patients, for which evidence of effectiveness was equivocal) but no further improvements were made. Possible explanations for this apparent stability are the "routinisation" of activities by staff and the higher expectations of patients, influenced by previous years' experiences. However, observed performance fell short of expectation in some cases, suggesting that withdrawing incentives is not without risk.

These findings should be interpreted in the light of the cost to payers of incentivising providers' performance, especially in the context of cost effectiveness and missed opportunities.[29] Modest but significant gains in performance are achievable in the first year or two for newly incentivised activities.[7 21 30] Although all the indicators we investigated were still indirectly or partially incentivised through other indicators that were not removed from the scheme, our findings indicate that withdrawing incentives for aspects of care for which performance has reached high levels and reinvesting in alternative aspects of care could provide an opportunity to drive improvement in the latter without greatly damaging quality of care in the former, thus maximising health benefits from incentive schemes. However, generalising the findings to all incentivised aspects of care would be premature, and careful

consideration needs to be given to aspects of care for which financial incentives are to be withdrawn.

This study is based on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this paper are those of the authors alone.

Contributors: EK and TD designed the study. DS extracted the data. EK and DR did the statistical analyses. EK and TD wrote the manuscript. DR, DS, JMV, and DA edited the manuscript. EK is the guarantor.

Funding: This study was funded by the National Institute for Health Research (NIHR) School for Primary Care Research, under the title "An investigation of the Quality and Outcomes Framework using the general practice research database" (project #141). This paper presents independent research funded by the NIHR. The views expressed are those of the authors 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/coi_disclosure.pdf (available on request from the corresponding author) and declare: EK was partly supported by an NIHR School for Primary Care Research fellowship in primary health care; TD was supported by an NIHR career development fellowship; 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: The study was approved by the independent scientific advisory committee (ISAC) for Clinical Practice Research Datalink research (reference number: 12_147Ra). No further ethics approval was required for the analysis of the data.

Transparency declaration: EK 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 have been explained.

Data sharing: Clinical Practice Research Datalink data cannot be shared owing to licensing restrictions.

1 Reeves D, Doran T, Valderas JM, Kontopantelis E, Trueman P, Sutton M, et al. How to identify when a performance indicator has run its course. *BMJ* 2010;340:c1717.
2 Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol* 2000;55:68-78.
3 Le Grand J. Motivation, agency, and public policy: of knights and knaves, pawns and queens. Oxford University Press, 2006.
4 May C, Finch T, Mair F, Ballini L, Dowrick C, Eccles M, et al. Understanding the implementation of complex interventions in health care: the normalization process model. *BMC Health Serv Res* 2007;7:1-7.
5 Lester H, Schmittdiel J, Selby J, Fireman B, Campbell S, Lee J, et al. The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators. *BMJ* 2010;340:c1898.
6 Roland M. Linking physicians' pay to the quality of care—a major experiment in the United Kingdom. *N Engl J Med* 2004;351:1448-54.
7 Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med* 2006;355:375-84.
8 Doran T, Kontopantelis E, Fullwood C, Lester H, Valderas JM, Campbell S. Exempting dissenting patients from pay for performance schemes: retrospective analysis of exception reporting in the UK Quality and Outcomes Framework. *BMJ* 2012;344:e2405.
9 Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. *N Engl J Med* 2008;359:274-84.
10 Bueving HJ, Bernsen RM, de Jongste JC, van Suijlekom-Smit LW, Rimmelzwaan GF, Osterhaus AD, et al. Influenza vaccination in children with asthma: randomized double-blind placebo-controlled trial. *Am J Respir Crit Care Med* 2004;169:488-93.
11 Cates CJ, Jefferson TO, Rowe BH. Vaccines for preventing influenza in people with asthma. *Cochrane Database Syst Rev* 2008;(2):CD000364.
12 Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open* 2013;3:e003190.
13 Kontopantelis E. A greedy algorithm for representative sampling: repsample in Stata. *J Stat Softw* 2013;56:1-18.
14 Communities and Local Government. The English indices of deprivation 2010: technical report. Department for Communities and Local Government, 2011.
15 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;372:728-36.
16 Doran T, Campbell S, Fullwood C, Kontopantelis E, Roland M. Performance of small general practices under the UK's Quality and Outcomes Framework. *Br J Gen Pract* 2010;60:e335-44.
17 Clinical codes repository. 2013. www.clinicalcodes.org/.
18 Kontopantelis E, Doran T, Gravelle H, Goudie R, Siciliani L, Sutton M. Family doctor responses to changes in incentives for influenza immunization under the U.K. Quality and Outcomes Framework pay-for-performance scheme. *Health Serv Res* 2012;47:1117-36.
19 National Health Service Information Centre. The Quality and Outcomes Framework. 2012. www.ic.nhs.uk/qof.
20 Kontopantelis E, Reeves D. metaan: random-effects meta-analysis. *Stata J* 2010;10:395-407.
21 Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *BMJ* 2011;342:d3590.
22 Collet D. Modelling binary data. Chapman and Hall, 1991.
23 Jain SH, Cassel CK. Societal perceptions of physicians: knights, knaves, or pawns? *JAMA* 2010;304:1009-10.
24 Deci EL, Koestner R, Ryan RM. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol Bull* 1999;125:627-68, discussion 92-700.
25 Benabou R, Tirole J. Intrinsic and extrinsic motivation. *Rev Econ Stud* 2003;70:489-520.
26 Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;380:37-43.
27 Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharm Ther* 2002;27:299-309.
28 Norbury M, Fawkes N, Guthrie B. Impact of the GP contract on inequalities associated with influenza immunisation: a retrospective population-database analysis. *Br J Gen Pract* 2011;61:e379-85.
29 Meacock R, Kristensen SR, Sutton M. The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. *Health Econ* 2013:1-13.
30 Kontopantelis E, Reeves D, Valderas JM, Campbell S, Doran T. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. *BMJ Qual Saf* 2013;22:53-64.

Accepted: 13 January 2014

Cite this as: *BMJ* 2014;348:g330

What is already known on this topic

The Quality and Outcomes Framework is a very expensive pay for performance programme that has shaped UK primary care since 2004

Under the scheme, increases in performance and a closing of the inequality gap have been observed

Improvements could be attributed to increasing performance trends before incentivisation, and the scheme might have led to a small neglect of non-incentivised aspects of care

The scheme is regularly reviewed, and indicators have been withdrawn from it, but the effect of this on levels of care is unknown

What this study adds

The removal of incentives—although only partially—seems to have had a very small effect on quality of care, even over the long term

As new incentives can lead to quick gains in quality of care, replacing existing indicators with little potential for further improvement could provide an opportunity to maximise health benefits from incentive schemes

Tables

Table 1| Practices' characteristics for main and sensitivity datasets

	All available and up to standard practices within each year*			All available and up to standard practices across whole time period (sensitivity 1) (2004/05 to 2011/12)	Sample of 50 practices nationally representative on list size and deprivation (sensitivity 2) (2004/05 to 2011/12)
Characteristics	**2004/05***	**2008/09***	**2011/12***		
Mean (SD) list size†	9152.7 (4702.2)	9643.7 (5211.9)	10157.8 (5579.8)	9111.9 (4535.6)	6597.4 (3391.5)
No of practices	554	565	499	452	50
No (%) of practices by national deprivation fifth:					
0 (most affluent)	92 (17)	96 (17)	87 (17)	83 (18)	10 (20)
1	109 (20)	112 (20)	105 (21)	95 (21)	10 (20)
2	120 (22)	122 (22)	107 (21)	98 (22)	10 (20)
3	124 (22)	128 (23)	103 (21)	88 (19)	10 (20)
4 (most deprived)	109 (20)	107 (19)	97 (19)	88 (19)	10 (20)
No (%) of practices by region‡:					
North East	10 (2)	11 (2)	8 (2)	7 (2)	0 (0)
North West	73 (13)	72 (13)	60 (12)	59 (13)	10 (20)
Yorkshire and the Humber	24 (4)	17 (3)	9 (2)	9 (2)	0 (0)
East Midlands	19 (3)	15 (3)	6 (1)	5 (1)	0 (0)
West Midlands	47 (8)	47 (8)	41 (8)	39 (9)	2 (4)
East of England	44 (8)	39 (7)	29 (6)	27 (6)	2 (4)
South West	47 (8)	50 (9)	44 (9)	38 (8)	2 (4)
South Central	50 (9)	53 (9)	53 (11)	49 (11)	3 (6)
London	61 (11)	71 (13)	71 (14)	55 (12)	7 (14)
South East Coast	51 (9)	54 (10)	51 (10)	46 (10)	7 (14)
Northern Ireland	21 (4)	21 (4)	20 (4)	19 (4)	4 (8)
Scotland	63 (11)	65 (12)	61 (12)	58 (13)	6 (12)
Wales	44 (8)	50 (9)	46 (9)	41 (9)	7 (14)

*Characteristics for representative three out of eight time points presented for main analysis.

†For 8486 English practices participating in Quality and Outcomes Framework (QOF) in 2004/05 (99.9% of all), mean list size was 6226 (SD 3869). For 8128 practices in 2011/12, mean list size was 6836 (SD 4274). For sensitivity samples 1 and 2, mean list size is practice average over all time points.

‡As reported by QOF, regional breakdown of English practices for 2011/12 was: North East=404 (5.0%), North West=1254 (15.4%), Yorkshire and the Humber=785 (9.7%), East Midlands=621 (7.6%), West Midlands=958 (11.8%), East of England=787 (9.7%), South West=719 (8.8%), South Central=501 (6.2%), London=1472 (18.1%), South East Coast=622 (7.7%).

Table 2| Removed indicators and their linked intermediate outcome, process, and condition indicators, by analysis

Analysis	Removed indicator	Description*	Type	Maximum points (remuneration)†	Active	Linked outcome indicator‡	Covariate indicators	
							Process related§	Condition related¶
Analyses performed								
Influenza immunisation**	Asthma7	Patients (aged 16 and over) with asthma who have had influenza immunisation	Treatment	12 (£1527)	2004/05-2005/06	—	CHD12, COPD8 DM18, Stroke10	Asthma5
Blood pressure	CHD5	Patients with coronary heart disease with a record of blood pressure	Monitoring	7 (£890)	2004/05-2010/11	CHD6	BP4††, DM11, Stroke5	CHD2/13
	DM11	Patients with diabetes with a record of blood pressure	Monitoring	3 (£382)	2004/05-2010/11	DM12/30	BP4††, CHD5, Stroke5	DM2, DM14/22
	Stroke5	Patients with transient ischaemic attack or stroke with a record of blood pressure	Monitoring	2 (£255)	2004/05-2010/11	Stroke6	BP4††, CHD5, DM11	Stroke7
Cholesterol	CHD7	Patients with CHD with a record of total cholesterol	Monitoring	7 (£890)	2004/05-2010/11	CHD8	Stroke7††, DM16	CHD2/13
	DM16	Patients with diabetes with a record of total cholesterol	Monitoring	3 (£382)	2004/05-2010/11	DM17	Stroke7††, CHD7	DM2, DM14/22
Blood glucose	DM5	Patients with diabetes with a record of HbA_{1c} or equivalent	Monitoring	3 (£382)	2004/5-2010/11	DM6/20/23/26	—	DM2, DM14/22
Lithium therapy**	MH3	Patients on lithium therapy with a record of lithium levels (previous 6 months)	Monitoring	3 (£382)	2004/05-2005/06	MH5/18	—	MH4/17
Analyses not performed‡‡								
Medication review in patients with epilepsy	EPI3/7	Patients (aged 18 and over) on drug treatment for epilepsy with record of medication review	Monitoring	4 (£509)	2004/05-2010/11	—	—	EPI2/6
Follow-up of patients with schizophrenia, bipolar affective disorder, or other psychoses	MH7	Patients with psychosis who do not attend for annual review who are followed up within 14 days	Monitoring	3 (£382)	2006/07-2010/11	—	—	MH4/17
Spirometry	COPD2	Patients with newly diagnosed COPD where diagnosis has been confirmed by spirometry and reversibility testing	Monitoring	5 (£636)	2004/05-2005/06	—	Asthma2/8	COPD3

CHD=coronary heart disease; COPD=chronic obstructive pulmonary disease.

*Indicator must be achieved in previous 15 months, unless otherwise specified.

†Maximum points available for indicator in year before removal. Points are converted into payments at rate of £127.26 per point (in 2011/12), adjusted for practice size and disease prevalence.

‡Some of removed monitoring or measurement indicators are "linked" to outcome indicators that remained in scheme; for example, blood pressure measurement is "linked" to blood pressure control.

§Process indicators from different disease domain for which process/action is similar or identical in removed indicator, selected to act as within process controls in analyses (for example, influenza immunisation in patients with asthma and patients with CHD).

¶Process indicators from same disease domain as removed indicator, selected to act as within condition controls in analyses (for example, influenza immunisation for asthma patients and smoking cessation advice for smokers with asthma).

**As explained in methods section, covariate indicators were not included in long term effects analyses (influenza immunisation; lithium therapy), but they are listed here for completeness.

††BP4 (blood pressure monitoring in hypertension patients) and Stroke7 (cholesterol monitoring in stroke patients) are linked with outcome indicators BP5 and Stroke8 respectively. BP5 and Stroke8 have been modelled in Clinical Practice Research Datalink (CPRD) for comparison with intermediate outcome indicators that are linked with removed indicators and are used as controls in outcome analyses.

Table 2 (continued)

Analysis	Removed indicator	Description*	Type	Maximum points (remuneration)†	Active	Linked outcome indicator‡	Covariate indicators	
							Process related§	Condition related¶

‡‡Analyses for indicators covering medication review in epilepsy (EPI3/7), follow-up of severe mental health disorders (MH7), and spirometry (COPD2) could not be done, as indicators were too complex to model in CPRD (MH7) or respective code lists failed to capture plausible trends and levels compared with those reported under Quality and Outcomes Framework (EPI3/7, COPD2).

Table 3| **Mean (SD) practice prevalence scores for main analysis dataset, compared with national scores.**

Condition	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12
All available and up to standard CPRD practices within each year—main analysis								
Asthma	5.74 (1.28)	5.67 (1.26)	5.58 (1.22)	5.43 (1.23)	5.42 (1.26)	5.35 (1.27)	5.11 (1.25)	4.82 (1.22)
Hypertension	11.38 (2.65)	11.74 (2.79)	12.00 (2.92)	12.11 (3.00)	12.22 (3.01)	12.18 (3.05)	12.21 (3.07)	12.05 (3.09)
CHD	3.59 (1.37)	3.48 (1.33)	3.40 (1.28)	3.29 (1.26)	3.19 (1.20)	3.07 (1.17)	2.97 (1.13)	2.84 (1.03)
COPD	1.11 (0.60)	1.17 (0.62)	1.22 (0.64)	1.27 (0.67)	1.31 (0.68)	1.33 (0.70)	1.38 (0.74)	1.39 (0.73)
Diabetes*	3.13 (0.75)	3.30 (0.81)	3.45 (0.87)	3.58 (0.92)	3.71 (0.95)	3.83 (1.02)	3.97 (1.07)	4.09 (1.14)
Epilepsy†	0.59 (0.19)	0.59 (0.19)	0.59 (0.19)	0.58 (0.19)	0.58 (0.19)	0.55 (0.18)	0.55 (0.19)	0.54 (0.19)
Mental health‡ (lithium therapy only)	0.13 (0.07)	0.12 (0.07)	0.11 (0.06)	0.11 (0.06)	0.10 (0.06)	0.10 (0.06)	0.10 (0.06)	0.10 (0.06)
Stroke	1.64 (0.50)	1.64 (0.50)	1.64 (0.50)	1.63 (0.50)	1.62 (0.49)	1.62 (0.50)	1.63 (0.50)	1.63 (0.51)
No of practices	554	567	569	566	565	556	534	499
All English practices, as reported by QOF								
Asthma	5.67 (1.63)	5.73 (1.50)	5.69 (1.42)	5.68 (1.39)	5.82 (1.45)	5.95 (2.29)	5.89 (1.42)	5.90 (1.63)
Hypertension	11.16 (3.64)	11.93 (3.52)	12.50 (3.51)	12.84 (3.53)	13.2 (3.70)	13.52 (4.44)	13.60 (3.84)	13.75 (3.94)
CHD	3.54 (1.42)	3.55 (1.36)	3.52 (1.31)	3.49 (1.28)	3.47 (1.42)	3.45 (1.49)	3.39 (1.38)	3.38 (1.37)
COPD	1.36 (0.86)	1.39 (0.82)	1.45 (0.80)	1.51 (0.81)	1.57 (0.87)	1.63 (0.96)	1.68 (0.92)	1.74 (0.91)
Diabetes*	3.42 (1.04)	3.64 (1.07)	4.98 (5.59)	4.96 (1.46)	5.26 (1.74)	5.53 (2.21)	5.76 (2.67)	6.00 (2.91)
Epilepsy†	0.58 (0.24)	0.60 (0.23)	1.07 (5.54)	0.76 (0.28)	0.76 (0.31)	0.77 (0.38)	0.82 (2.21)	0.83 (2.23)
Mental health‡	0.57 (0.52)	0.63 (0.64)	0.74 (0.47)	0.77 (0.50)	0.79 (0.55)	0.84 (0.80)	0.84 (0.61)	0.87 (0.62)
Stroke	1.40 (0.78)	1.50 (0.75)	1.55 (0.75)	1.58 (0.74)	1.62 (0.91)	1.65 (0.97)	1.67 (0.87)	1.70 (0.86)
No of practices	8486	8406	8372	8294	8229	8305	8245	8123

CHD=coronary heart disease; COPD=chronic obstructive pulmonary disease; CPRD=Clinical Practice Research Datalink; QOF=Quality and Outcomes Framework.

*For patients aged 17 or over (except for QOF prevalence scores in 2004/05 and 2005/06, when no age restriction was applied).

†For patients aged 18 or over (except for QOF prevalence scores in 2004/05 and 2005/06, when no age restriction was applied).

‡Mental health as reported in QOF relates to all diagnoses of psychosis, irrespective of treatment, whereas in our analyses we focused on patients treated with lithium.

Table 4| **Observed mean (SD) practice indicator scores (percentage achievement rates) over time, by group**

Indicator	Description	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12
Influenza immunisation									
Asthma7*	Patients with asthma immunised against influenza	78.0 (7.0)	78.2 (6.8)	78.0 (6.9)†	78.2 (6.9)†	78.0 (6.9)†	78.2 (6.6)†	79.0 (6.6)†	78.8 (6.5)†
CHD12	Patients with CHD immunised against influenza	90.5 (4.8)	90.7 (4.7)	90.8 (4.7)	90.7 (4.8)	90.5 (4.8)	90.4 (4.8)	90.2 (4.5)	89.5 (4.5)
COPD8	Patients with COPD immunised against influenza	92.1 (5.2)	92.1 (5.3)	92.3 (4.8)	92.2 (4.7)	92.1 (4.6)	91.9 (4.4)	91.2 (4.5)	90.1 (4.6)
DM18	Patients with diabetes immunised against influenza	90.1 (4.8)	90.2 (4.3)	90.1 (4.3)	89.8 (4.4)	89.5 (4.4)	89.1 (4.4)	88.5 (4.3)	87.2 (4.3)
Stroke10	Patients with stroke immunised against influenza	88.4 (4.9)	88.7 (4.9)	88.6 (4.8)	88.5 (4.9)	88.2 (5.1)	88.0 (5.2)	87.6 (5.1)	86.6 (5.1)
Blood pressure									
CHD5‡	Patients with CHD with record of blood pressure	93.6 (5.0)	94.0 (4.3)	94.2 (4.1)	94.2 (4.1)	93.9 (4.2)	93.8 (4.1)	94.1 (3.9)	93.9 (3.9)†
DM11‡	Patients with diabetes with record of blood pressure	94.0 (4.1)	94.3 (3.6)	94.4 (3.1)	94.2 (3.0)	94.1 (3.3)	94.0 (3.3)	94.3 (3.0)	94.3 (3.0)†
Stroke5‡	Patients with stroke with record of blood pressure	91.4 (5.3)	92.0 (6.2)	92.5 (4.3)	92.7 (4.1)	92.5 (4.4)	92.5 (4.5)	92.7 (4.4)	92.4 (4.5)†
BP4	Patients with hypertension with record of blood pressure	88.5 (5.1)	90.0 (5.6)	90.4 (3.8)	90.1 (3.6)	89.5 (3.8)	89.4 (3.6)	89.7 (3.5)	89.6 (3.4)
CHD6	Patients with CHD, last blood pressure ≤150/90 mm Hg	84.6 (6.6)	86.1 (5.7)	87.8 (5.5)	88.4 (5.2)	89.1 (4.9)	89.5 (5.0)	90.0 (4.6)	90.5 (4.5)
DM12/30	Patients with diabetes, last blood pressure ≤145/85 mm Hg	68.3 (9.9)	70.4 (9.2)	73.8 (8.6)	75.0 (8.6)	75.8 (8.3)	76.5 (8.2)	77.8 (7.9)	78.2 (7.5)
Stroke6	Patients with stroke, last blood pressure ≤150/90 mm Hg	81.8 (7.3)	83.7 (6.7)	85.9 (6.0)	86.7 (6.1)	87.4 (5.7)	88.1 (5.6)	88.8 (5.1)	89.0 (5.0)
BP5	Patients with hypertension, last blood pressure ≤150/90 mm Hg	75.1 (8.4)	77.7 (7.2)	79.9 (6.8)	81.1 (6.5)	81.9 (6.2)	82.9 (6.1)	83.8 (5.9)	84.6 (5.6)
Cholesterol									
CHD7‡	Patients with CHD with record of total cholesterol	85.2 (9.4)	87.2 (8.1)	88.2 (6.6)	88.4 (6.3)	88.1 (6.3)	88.1 (6.2)	88.3 (5.9)	87.0 (6.6)†
DM16‡	Patients with CHD with record of total cholesterol	89.7 (6.1)	90.8 (5.8)	91.2 (4.1)	91.1 (3.9)	91.1 (4.1)	91.1 (3.8)	91.4 (3.6)	91.2 (3.8)†
Stroke7	Patients with CHD with record of total cholesterol	77.6 (11.0)	81.8 (10.1)	83.9 (7.6)	84.7 (6.9)	84.6 (6.8)	85.0 (7.1)	85.3 (7.1)	85.5 (6.9)
CHD8	Patients with CHD, last total cholesterol ≤5 mmol/L	78.4 (8.1)	82.2 (6.5)	84.5 (5.7)	85.1 (5.2)	84.8 (5.1)	84.7 (5.1)	84.8 (4.7)	84.4 (4.8)
DM17	Patients with diabetes, last total cholesterol ≤5 mmol/L	78.6 (8.2)	83.1 (6.1)	85.3 (5.3)	85.6 (4.9)	85.1 (4.8)	84.9 (4.6)	84.5 (4.4)	83.6 (4.6)
Stroke8	Patients with stroke, last total cholesterol ≤5 mmol/L	72.5 (8.4)	77.5 (7.7)	80.7 (6.7)	81.5 (6.1)	81.2 (6.3)	81.4 (5.7)	81.5 (5.5)	81.1 (5.5)
Glycaemic control									
DM5‡	Patients with diabetes with record of HbA_{1c}	90.5 (5.6)	91.2 (5.6)	91.3 (4.1)	91.2 (3.9)	91.3 (4.0)	91.6 (3.8)	92.1 (3.6)	92.2 (3.7)†
DM6/20/23/26	Patients with diabetes, last HbA_{1c} ≤7.5§	64.8 (9.9)	66.5 (9.3)	68.7 (9.0)	68.7 (9.1)	69.5 (8.1)	70.7 (7.3)	71.4 (7.0)	70.4 (7.2)
Lithium therapy									
MH3*	Patients on lithium therapy with a record of lithium levels	89.2 (14.3)	91.1 (14.0)	91.4 (12.0)†	91.0 (12.9)†	91.8 (12.8)†	92.1 (11.4)†	92.4 (11.1)†	92.5 (12.4)†
MH5/18	Patients on lithium therapy with lithium levels in therapeutic range	86.7 (22.8)	87.9 (20.4)	89.0 (19.8)	90.7 (17.1)	91.5 (16.6)	90.9 (16.9)	90.5 (17.6)	89.6 (18.0)

Table 4 (continued)

Indicator	Description	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12
Within condition control indicators									
Asthma5 (Smoking2/4)	Patients with asthma who smoke offered smoking cessation advice	64.3 (10.9)	66.6 (10.4)	72.0 (9.3)	74.0 (8.6)	77.0 (8.9)	79.5 (8.1)	87.8 (5.7)	88.6 (5.6)
CHD2/13	Patients with newly diagnosed angina referred for specialist assessment	61.1 (18.0)	70.0 (17.1)	74.5 (15.3)	76.7 (15.3)	79.6 (14.2)	81.4 (14.2)	80.4 (15.1)	80.4 (15.1)
BP2 (Smoking1/3)	Patients with hypertension with a record of smoking status	76.4 (14.5)	75.5 (14.3)	74.8 (14.1)	74.2 (13.8)	73.4 (13.7)	72.8 (13.5)	72.1 (13.3)	71.5 (13.1)
DM2	Patients with diabetes with a record of body mass index	88.4 (6.9)	89.6 (5.4)	90.2 (4.7)	89.9 (4.5)	89.7 (4.7)	89.4 (4.8)	89.6 (4.6)	90.0 (4.4)
DM14/22	Patients with diabetes with a record of serum creatinine	90.2 (6.3)	91.5 (5.9)	92.0 (4.2)	91.9 (4.1)	91.9 (5.0)	91.8 (6.1)	92.5 (5.8)	92.7 (5.2)
MH4/17	Patients on lithium therapy with a record of serum creatinine and thyroid stimulating hormone	87.9 (17.3)	92.3 (12.2)	92.6 (12.3)	93.5 (10.2)	94.1 (10.6)	95.0 (8.9)	94.6 (9.8)	95.6 (8.9)
Epilepsy2/6	Patients on drug treatment for epilepsy with a record of seizure frequency	87.6 (12.5)	89.1 (8.6)	89.4 (7.7)	89.8 (7.6)	89.8 (7.4)	89.1 (7.6)	89.7 (7.4)	90.0 (8.2)

CHD=coronary heart disease; COPD=chronic obstructive pulmonary disease.

*Removed from April 2006.

†Post-incentivisation rates.

‡Removed from April 2011.

§Required value for HbA_{1c} regulation indicator changed over time: 7.4% in 2004/05 and 2005/06; 7.5% in 2006/07 to 2008/9; 7.0% in 2009/10 and 2010/11; 7.5% in 2011/12 and 2012/13. For consistency, 7.5% was set as target for glycaemic regulation.

Table 5| **Short term effects—mean back transformed observed and predicted scores and their difference (95% CI)***

Indicator	Measure	2008/09†	2009/10†	2010/11†	2011/12‡
Removed indicators§					
CHD5	Observed (predicted)	95.0 (94.9)	94.8 (94.9)	95.1 (95.0)	95.0 (95.0)
	Difference (95% CI)	0.08 (−0.12 to 0.28)	−0.15 (−0.33 to 0.04)	0.08 (−0.11 to 0.27)	−0.02 (−0.24 to 0.20)
DM11	Observed (predicted)	94.9 (94.8)	94.8 (94.9)	95.0 (95.0)	95.0 (95.0)
	Difference (95% CI)	0.04 (−0.10 to 0.18)	−0.08 (−0.20 to 0.04)	0.04 (−0.08 to 0.17)	0.04 (−0.10 to 0.18)
Stroke5	Observed (predicted)	93.6 (93.6)	93.6 (93.6)	93.7 (93.7)	93.5 (93.7)
	Difference (95% CI)	0.01 (−0.18 to 0.21)	−0.01 (−0.23 to 0.21)	0.01 (−0.20 to 0.22)	−0.18 (−0.43 to 0.07)
CHD7	Observed (predicted)	89.3 (89.3)	89.3 (89.3)	89.5 (89.5)	88.4 (89.6)
	Difference (95% CI)	0.01 (−0.30 to 0.32)	−0.003 (−0.33 to 0.33)	0.01 (−0.30 to 0.32)	−1.19 (−1.56 to −0.81)
DM16	Observed (predicted)	91.9 (91.8)	91.8 (91.9)	92.1 (92.1)	92.0 (92.2)
	Difference (95% CI)	0.04 (−0.14 to 0.21)	−0.07 (−0.22 to 0.08)	0.04 (−0.12 to 0.20)	−0.18 (−0.36 to 0.001)
DM5	Observed (predicted)	92.0 (92.0)	92.4 (92.4)	92.8 (92.8)	93.0 (93.2)
	Difference (95% CI)	0.02 (−0.14 to 0.19)	−0.04 (−0.19 to 0.11)	0.02 (−0.13 to 0.17)	−0.15 (−0.32 to 0.02)
Linked outcome indicators					
CHD6	Observed (predicted)	90.1 (90.1)	90.6 (90.6)	91.0 (91.0)	91.4 (91.3)
	Difference (95% CI)	−0.02 (−0.25 to 0.21)	0.03 (−0.19 to 0.25)	−0.02 (−0.23 to 0.20)	0.14 (−0.11 to 0.39)
DM12/30	Observed (predicted)	76.8 (76.7)	77.6 (77.8)	78.9 (78.8)	79.2 (79.5)
	Difference (95% CI)	0.12 (−0.29 to 0.54)	−0.23 (−0.62 to 0.16)	0.12 (−0.27 to 0.52)	−0.28 (−0.72 to 0.17)
Stroke6	Observed (predicted)	88.6 (88.6)	89.4 (89.3)	89.9 (90.0)	90.1 (90.4)
	Difference (95% CI)	−0.04 (−0.34 to 0.26)	0.09 (−0.20 to 0.38)	−0.04 (−0.32 to 0.24)	−0.35 (−0.65 to −0.05)
CHD8	Observed (predicted)	85.5 (85.5)	85.4 (85.4)	85.4 (85.4)	85.0 (85.3)
	Difference (95% CI)	0.03 (−0.22 to 0.28)	−0.04 (−0.30 to 0.23)	0.03 (−0.23 to 0.29)	−0.32 (−0.62 to −0.02)
DM17	Observed (predicted)	85.8 (85.8)	85.5 (85.4)	85.0 (85.1)	84.2 (84.6)
	Difference (95% CI)	−0.03 (−0.28 to 0.22)	0.04 (−0.19 to 0.28)	−0.02 (−0.26 to 0.23)	−0.45 (−0.75 to −0.15)
DM6/20/23/26	Observed (predicted)	69.9 (70.0)	71.3 (71.0)	72.0 (72.1)	71.0 (73.0)
	Difference (95% CI)	−0.13 (−0.54 to 0.29)	0.24 (0.01 to 0.47)	−0.12 (−0.39 to 0.15)	−2.08 (−2.45 to −1.71)

*For indicators for which denominators are small and 100% scores are prevalent, discrepancies can exist between true and back transformed scores owing to empirical logit (that is, score at 100% is back transformed to lower score).

†Pre-removal time points used in modelling; predictions for 2008/09-2010/11 indicate good fit of linear models used.

‡Post-removal time points.

§All indicators removed in April 2011.

Figures

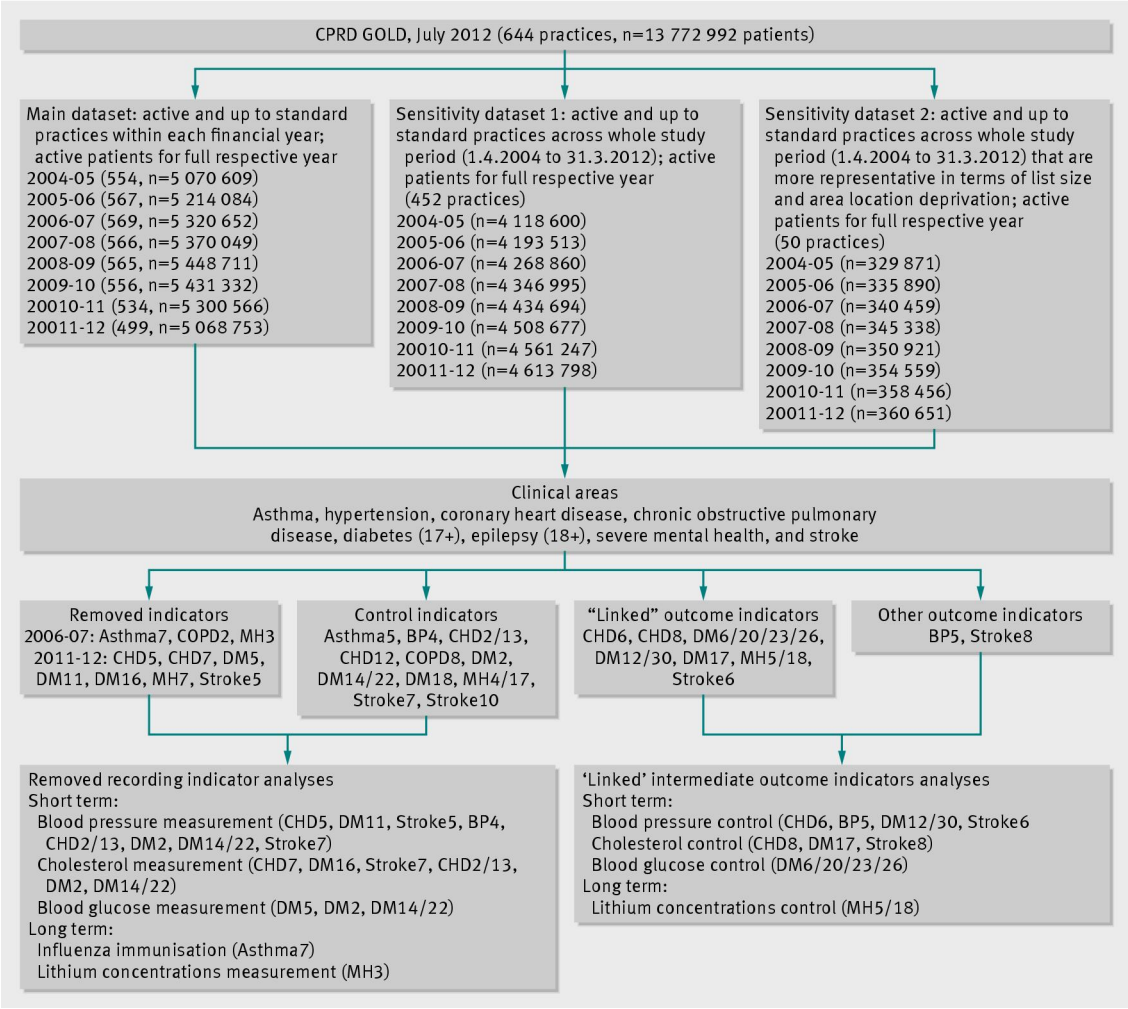

Fig 1 Flow chart of dataset creation and analyses. Only successfully modelled indicators are listed. Indicator details are provided in tables 2 and 4 and in web appendix table A1

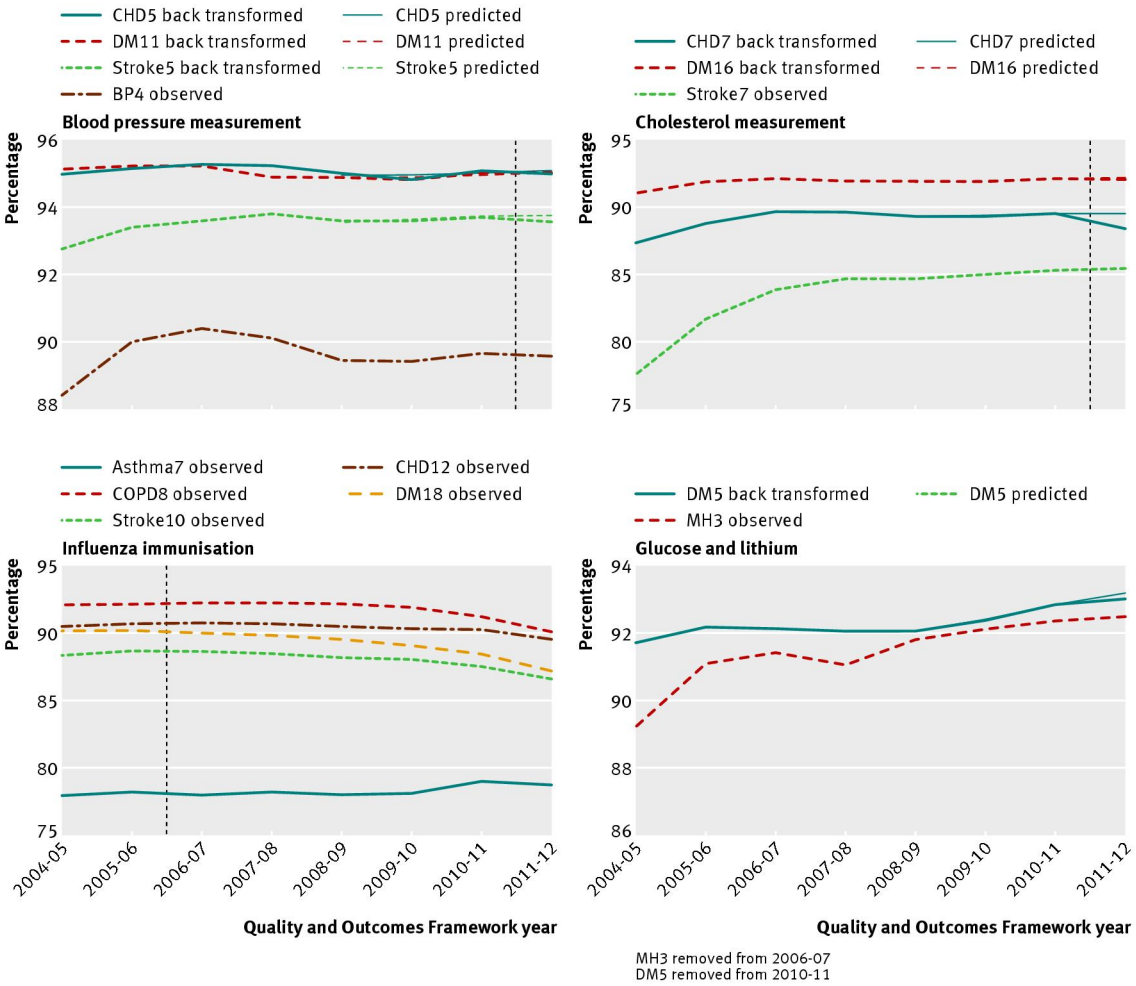

Fig 2 Trends and predictions for removed and related unremoved indicators. For indicators removed in April 2011, predicted scores were compared with back transformed observed scores (from logit). Although back transformed observed scores agree with raw scores fully in most cases, that might not be true for indicators for which denominators are small and 100% scores are prevalent. This can lead to discrepancies due to empirical logit (that is, score at 100% is back transformed to lower score) and an "unfair" comparison between observed and predicted. Unremoved process related control indicators were also plotted (using raw scores as no comparison with predictions exists). Condition related control indicators were not plotted; vertical lines indicate timing of indicator removal

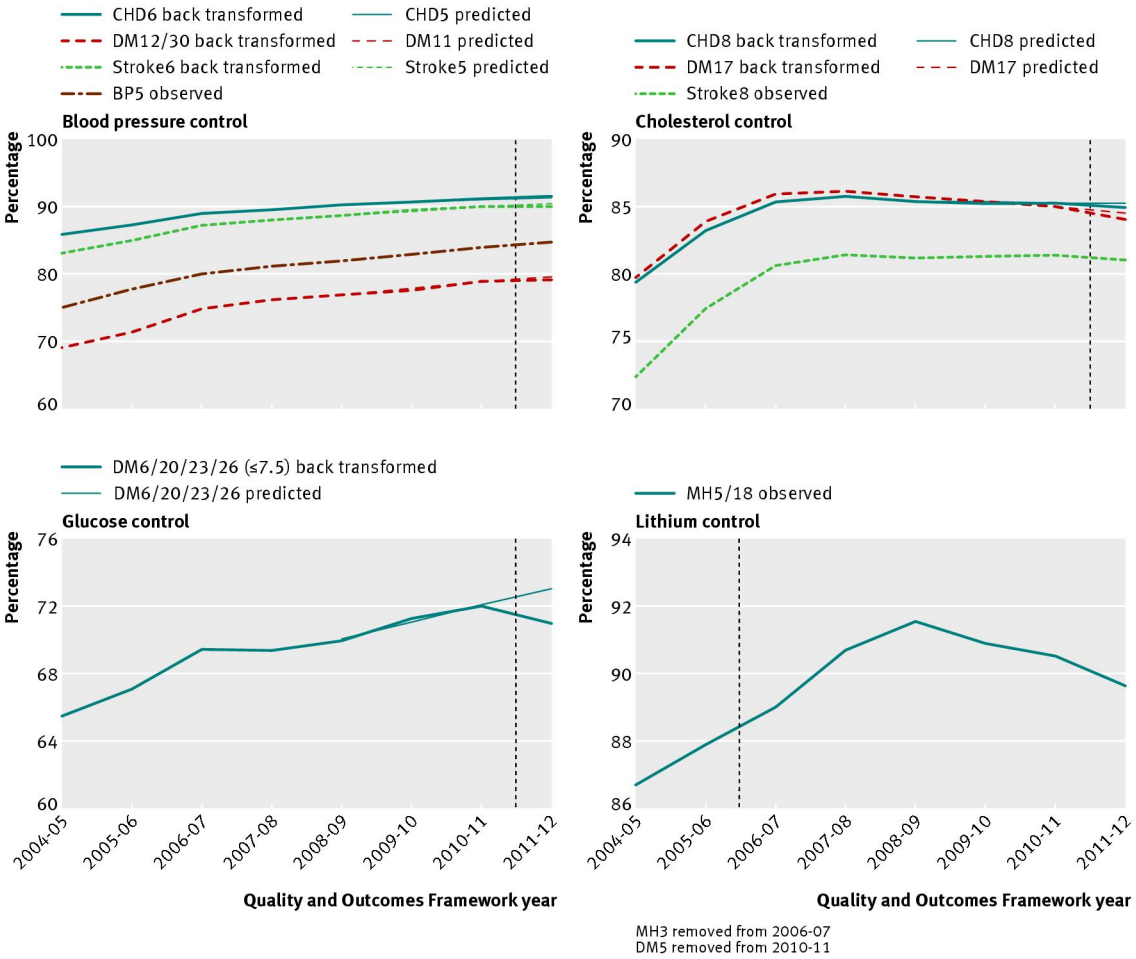

Fig 3 Trends and predictions for "linked" unremoved outcome indicators and related indicators. For short term removal effects on the linked outcome indicators, predicted scores were compared with back transformed observed scores (from logit). Although back transformed observed scores agree with raw scores fully in most cases, that might not be true for indicators for which denominators are small and 100% scores are prevalent. This can lead to discrepancies due to empirical logit (that is, score at 100% is back transformed to lower score) and an "unfair" comparison between observed and predicted. Unremoved process related control indicators were also plotted (using raw scores as no comparison with predictions exists). Vertical lines indicate timing of "linked" process indicator removal