Accounting for the Quality of NHS Output

Chris Bojke, Adriana Castelli,
Katja Grašič, Anne Mason,
Andrew Street

Accounting for the quality of NHS output

[2]Chris Bojke
[1]Adriana Castelli
[1]Katja Grašič
[1]Anne Mason
[3]Andrew Street

[1]Centre for Health Economics, University of York, York, UK
[2]Leeds Institute of Health Sciences, University of Leeds, UK
[3]Department of Health Policy, The London School of Economics and Political Science, UK

April 2018

Background to series

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Further copies

Only the latest electronic copy of our reports should be cited. Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/ Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321405 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/che

Abstract

Output measures used in the national accounts aspire to capture as comprehensively and accurately as possible the value that society places on everything produced by the economy. Given that economies produce heterogeneous products, some means of defining and valuing these is required so that a single aggregate measure of output can be constructed.

For products traded in the market economy this is conceptually quite straightforward, but it requires the assumption that prices reflect marginal social values and equate to the marginal costs of production. For products and services made available by the 'non-market' economy, encompassing sectors such as defence, education and health systems, among others, the above assumption does not hold. People access and use the services provided by these sectors but rarely pay for them at point of use or, if they have to pay something out-of-pocket, it is usually subsidised.

So, for 'non-market' products, two ways have been proposed to construct an equivalent output measure: (1) to substitute information about the price of the output with its cost of production, making the assumption that marginal costs equate to marginal social values and (2) to describe and capture the characteristics of each product, recognising that products with more desirable characteristics are of greater value. In common parlance, this bundle of characteristics reflects the overall 'quality' of the product.

A combination of these two general approaches has been adopted to assess the contribution of the English National Health Service (NHS) in the national accounts. Current practice in accounting for the quality of healthcare services makes use of routinely available information in order to capture the QALYs associated with treating patients, by combining information on survival rates, life expectancy and a measure of change in health status before and after treatment. The process of care delivery is captured by measures of treatment waiting times.

This approach may overlook other important characteristics of the quality of healthcare. This review provides the conceptual framework needed to select potentially appropriate characteristics of healthcare outputs. To this end we evaluated three published sets of criteria developed by national bodies responsible for assessing healthcare system performance. We also sought the opinions of UK experts on quality expressed at a workshop. From this process seven criteria were established. We next reviewed two sources of quality indicators currently collected and reported for the English NHS: the NHS Outcomes Framework indicators and NHS Thermometer indicators. A schema, including indicator name and source, data source, time period covered, definitions and purpose, was developed for each of the indicators. Indicators were individually assessed by the research team, and one expert from the Department of Health and one from the Office for National Statistics in order to establish whether they met each of the identified criteria. Depending on the level of consensus among reviewers, a maximum of 17 indicators were short-listed for potential use as quality adjustors for NHS output, all of which are NHS Outcomes Framework indicators.

Contents

1. Introduction

Output measures used in the national accounts aspire to capture as comprehensively and accurately as possible the value that society places on everything produced by the economy. Given that economies produce heterogeneous products, some means of defining and valuing these is required so that a single aggregate measure of output can be constructed.

For products traded in the market economy (*Mkt*) this is conceptually quite straightforward, but it requires the assumption that prices reflect marginal social values and equate to the marginal costs of production. If this is accepted, output measurement reduces to a collection of information about the volume (x) and price (p) of each traded product($j = 1 \dots J$). We can combine these to form the following aggregate measure of output (Y) for the sector (s) in question:

$$Y_s^{Mkt} = \sum_j x_j p_j \tag{1.1}$$

To measure growth in output, the volumes of each product are compared across consecutive periods, holding prices constant. We can use either prices from the current or the base period. If using prices from the base period (p_{jt-1}) the Laspeyres index (*L*) of output growth is specified as:

$$\Delta Y_{s\_L}^{Mkt} = \frac{\sum_j x_{jt} p_{jt-1}}{\sum_j x_{jt-1} p_{jt-1}} \tag{1.2}$$

If current prices (p_{jt}) are used, the Paasche index (*P*) is specified as:

$$\Delta Y_{s\_P}^{Mkt} = \frac{\sum_j x_{jt} p_{jt}}{\sum_j x_{jt-1} p_{jt}} \tag{1.3}$$

There is a slight difference in the interpretation between the two indices. In the case of Laspeyres index, the individual can afford the same basket of products in the current period as in the base period. Conversely, with the Paasche index the assumption is that the individual could have afforded the same goods in the previous period as she can now.

While these indices capture well the output in the market economy, there are many things produced by the economy for which consumers do not have to pay the full price. The 'non-market' economy (*NMkt*) encompasses those sectors which are funded, wholly or partially, through taxation. In most countries these typically include government, and the justice, police, defence, education and health systems, among others [1]. People access and use the services provided by these sectors but rarely pay for these services at point of use or, if they have to pay something out-of-pocket, it is usually subsidised.

This means that the assumption we made about products traded in the market economy - *that prices reflect marginal social values and equate to the marginal costs of production* – does not hold. While it may be possible to collect information about the volume of services provided, information on prices is unavailable. As a consequence, output measurement for non-market sectors is less straightforward than for market sectors.

There are two ways to overcome the problem and construct an equivalent output measure for non-market. The first way is to substitute information about the price of the output with its cost of production, making the assumption that marginal costs equate to marginal social values. If so, the output measure in Laspeyres form becomes:

$$\Delta Y_{s\_L}^{NMkt} = \frac{\sum_j x_{jt} c_{jt-1}}{\sum_j x_{jt-1} c_{jt-1}} \tag{1.4}$$

However, if marginal costs diverge from marginal social values, this index reflects producer rather than consumer valuations of output [2].

The second way requires a means of assessing the value of non-traded products. A common means of doing this is by describing the characteristics (g) of each product, recognising that products with more desirable characteristics are of greater value. The approach requires quantification of the various characteristics (q_g) of each product and assessing the marginal social value (π_g) of each characteristic. This makes it possible to construct an alternative output measure, whereby prices are replaced by a measure capturing the relative value of each product's characteristics [3], such that:

$$\Delta Y_{s\_L}^{NMkt} = \frac{\sum_j x_{jt} \sum_g q_{gjt} \pi_{gt-1}}{\sum_j x_{jt-1} \sum_g q_{gjt-1} \pi_{gt-1}} \tag{1.5}$$

In common parlance, this bundle of characteristics reflects the overall 'quality' of the product. Hence, construction of this measure requires assessment of the quality characteristics of each product.

A combination of these two general approaches has been adopted to assess the contribution of the English National Health Service (NHS) in the national accounts. In section 2 we describe the current approach used to capture changes in the costs and characteristics of healthcare outputs and the data used to measure these characteristics.

There are concerns, though, that other important characteristics are not captured adequately and that NHS output should account for additional indicators of the quality of healthcare [4]. In section 3 we consider criteria for selecting potentially appropriate characteristics of healthcare outputs. We first describe existing sets of criteria, focusing on those developed by national bodies responsible for assessing healthcare system performance. We also held a workshop to gather the opinions of UK experts on quality and productivity measurement.

In section 4 we set out seven criteria that indicators of the quality of health care services ought to satisfy in order to be considered as candidates for inclusion in a measure of NHS output growth.

In section 5, we assess the indicators published as part of the NHS Outcomes Framework and the NHS Thermometer data against the criteria set out in section 4. For this process we identify those indicators that offer the greatest potential to be included in the NHS output measure. We conclude in section 6.

2. Accounting for the quality of healthcare output

There is a great deal of variation among health service users in terms of the nature of their contact with the health system and what this contact seeks to achieve. To capture output, it is necessary to define and measure 'completed treatments', and this implies a time-limited unit of measurement. However, this is challenging particularly for patients with chronic conditions whose contact with the heath system is ongoing. Standard practice, therefore, has been to count the number of discrete activities (actions) undertaken by the various organisations that comprise the health sector [5].

Quality adjustment of these activities is difficult mainly because people do not demand healthcare for its own sake, but because of the contribution it makes to their health status. This requires some means of measuring the health outcome associated with treatment. People also value the process by which healthcare is delivered, such as whether they are treated with dignity and respect, and how quickly they can access services. Therefore, a measure of health care output should seek to capture aspects of both process and outcome of healthcare activities.

An obvious way of capturing the impacts of NHS treatment on health outcomes is to measure Quality Adjusted Life Years (QALYs). Therefore, this section first sets out how QALYs could be used in an ideal world, and considers the operational challenges of implementation in the real world (subsection 2.1). Given the absence of routinely available data on QALYs, the next subsection describes how quality adjustment is currently implemented using available data (2.2). Finally, we discuss a potential source of QALY data and consider its relevance and applicability for an alternative approach to the capturing the quality of NHS output.

2.1 Measuring health outcomes: the QALY approach

Ideally, measures of health outcome should indicate the value added to health as a result of contact with the health system. In the UK, a common metric to describe health outcome is the QALY, which captures information about both the length and quality of life. This can be used to assess the contribution of treatment to health outcomes, and is the metric recommended by the National Institute for Health and Care Excellence (NICE) in health technology assessment [6].

To see how QALYs are measured, consider a patient requiring an urgent heart operation (Figure 1), with life expectancy on the *x* axis and health-related quality of life on the *y* axis, with values ranging from 0=death to 1=perfect quality of life. Left untreated, she is expected to live for just one year with a poor quality of life (QoL = 0.4). However, if the patient receives treatment, she is expected to live for 5 years with a higher quality of life for each of these years (QoL = 0.6). Without treatment, the number of QALYs the patient is expected to have equals to *1*0.4 = 0.4* QALYs. If the operation takes place, the number of QALYs increases to *5*0.6 = 3 QALYs*. The *QALY* gain the patients enjoys when receiving the treatment is, therefore, *3-0.4=2.6 QALYs*.
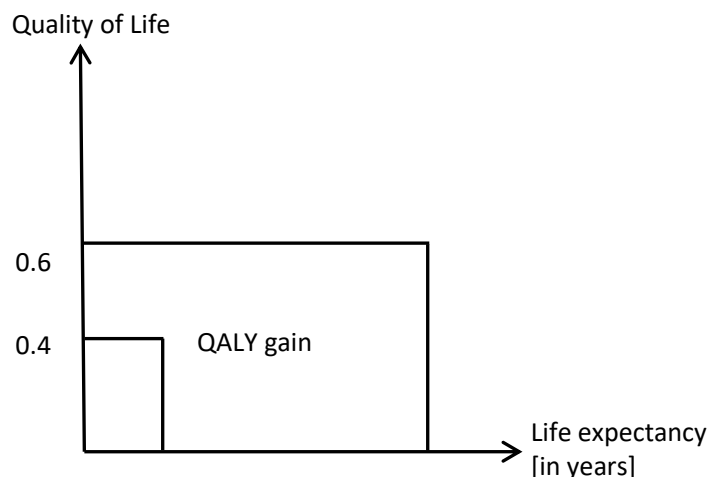
Figure 1: QALY gain

The QALY gain is the health produced by the healthcare system or NHS. If we could observe the health gains from all patients treated by the NHS over time, we could use this information to measure the performance of the NHS health system, using the total amount of QALYs to capture the total amount of health output produced by the health system.

So, if the without and with treatment number of QALYs is known for all patients receiving treatment of type *j* we can construct an output growth measure ΔY^{QALY} that measures the growth in total QALYs between years *t-1* and *t* aggregated across each activity *j* for the whole healthcare system:

$$\Delta Y^{QALY} = \frac{\sum_{i=1}^{N_t} v_{it}}{\sum_{i=1}^{N_{t-1}} v_{it-1}} \tag{2.1}$$

where v_{it} is the difference in *QALYs* without and with treatment (ie the number of *QALYs* gained) for patient *i* in year *t* and N_t is the total number of patients treated in year *t.*

There are three key challenges with operationalising this approach. First, there is the problem of attribution. Some improvements in health status may be due not to the activities of the health system, but reflect the influence of other types of care (e.g. social care) or of wider social determinants of health [7]. The challenge is to isolate the specific contribution of health services to health outcome.

Second, the without treatment counterfactual – what health status would have been in the absence of intervention – is rarely observed. Instead, health status measurement tends to rely on comparisons of health states before and sometime after intervention. For the purposes of measuring output growth in the national accounts, before and after measures can supply sufficient information on which to make temporal comparisons [7]. This would be the case if the counterfactual without-treatment profile can be assumed not to change from one year to the next. If so, before-and-after measures can be used to assess whether the with-treatment health profile changes over time, thus providing enough information with which to judge whether health outcomes have improved.

Third, data are not routinely collected about the health consequences of patients' contact with the health system. This makes it difficult to assess the impact of treatment on their quality of life. This lack of information is the reason why current practice in England has been to try to piece together measures of QALYs indirectly from other information. We describe this practice next.

2.2 The current approach to quality adjustment

In the absence of comprehensive and routinely collected data on QALYs, the current quality adjustment of NHS output makes use of routinely available information in order to capture the quality of life and extensions to length of life associated with treatments [3]. For patients treated in hospital, the adjustment takes the form:

$$\Delta q_j^{hosp} = \left(\frac{a_{jt}-k_j}{a_{jt-1}-k_j}\right) \frac{\left[\frac{\left(1-e^{-r_L L_{jt}}\right)}{r_L} - \frac{\left(e^{r_w w_{jt-1}}\right)}{r_w}\right]}{\left[\frac{\left(1-e^{-r_L L_{jt-1}}\right)}{r_L} - \frac{\left(e^{r_w w_{jt-1-1}}\right)}{r_w}\right]} \tag{2.2}$$

Given that direct QALY estimates for each type of hospital activity (defined using Healthcare Resource Groups (HRGs)) are unavailable, an equivalent of a QALY profile for patients allocated to each healthcare output is constructed [8]. A survival measure (a_j) captures the probability of survival following hospital treatment for people in each relevant HRG. We multiply this probability by life expectancy (LE_j) and a measure of change in health status following treatment (k_j) to arrive at an estimate of the total amount of QALYs experienced by this group of survivors over their remaining lifetime. Those who do not survive hospital treatment are afforded a zero QALY gain.

There is also recognition that the process of care delivery matters. Waiting for treatment (w_j) yields disutility, and this disutility is expressed in terms of QALYs by valuing days spent waiting in the same metric as the valuation of remaining life expectancy. This allows one to subtract the disutility associated with waiting from the QALY gains associated with treatment in order to arrive at an estimate of net QALY gains for each HRG.

Survival (a_j) is measured as the 30-day post admission survival rates for each output in each hospital. The change in health status (k_j) is measured as the ratio of average health status (h^0) before and after (h^*) treatment, such that $k_j = {h_j^0}/{h_j^*}$. In the absence of HRG-specific information, this ratio is assumed to be 0.8 for electives and 0.4 for non-electives and both remain constant over time [3].

For a handful of conditions, HRG-specific information about before and after treatment health status data is available via the Patient Reported Outcome Measure (PROM), which is a patient level survey asking about health status [9]. This survey is currently administered only to patients having hip replacement, knee replacement, hernia repair and varicose vein removal, representing less than 2% of all hospital patients. For patients having these treatments the change in health status, k_j, is taken from their pre- and post-treatment survey responses.

Life expectancy (LE_j) associated with each HRG is calculated by considering the age and gender profiles (in 5-year bands) of patients allocated to each HRG, based on life tables published by the Office of National Statistics (ONS) [10]. The inverse exponential function reflects decreasing life expectancy over time and r_Q is the discount rate applied to future life years.

Waiting times (w_j) for each HRG in each hospital are measured at the 80th percentile of the distribution for patients categorised to each HRG. This formulation implies that delays to treatment have adverse health consequences and that the marginal disutility of waiting increases as the delay increases, with the disutility captured as an exponential function and by the discount rate r_w [3].

The way that each type of hospital output is weighted to take account of its quality can be seen as a function of three ratios:

A survival and health effect ratio

$$\left(\frac{a_{jt} - k_j}{a_{jt-1} - k_j} \right)$$

A life expectancy ratio

$$\left(\frac{\left(1 - e^{-r_L L_{jt}}\right)}{\left(1 - e^{-r_L L_{jt-1}}\right)} \right)$$

A waiting time ratio

$$\left(\frac{e^{-r_w w_{jt}} - 1}{e^{-r_w w_{jt-1}} - 1} \right)$$

Not all ratios are applied to each type of hospital activity. For example, patients treated as emergencies (non-elective patients) do not wait for treatment, so the waiting time ratio is assumed to be equal to 1. For outpatient activity, no survival and life expectancy data are available, and thus only the waiting time ratio is applied.

There is a different way of accounting for the quality of primary care [11]. The approach utilises data captured as part of the Quality and Outcomes Framework (QOF), under which GPs are rewarded for achieving a range of diverse targets. If disease management in primary care is improving over time, the supposition is that this will be reflected in reduced blood pressure for an increasing proportion of patients with coronary artery disease (CHD), stroke and hypertension. Hence, primary care consultations are deemed to be 30% more valuable if a blood pressure reading equal or below the target of 150/90 is recorded.

To incorporate these aspects of quality into an output index for primary care, information is required about the prevalence rate for each of these three conditions, the QOF success rate, and the value of a consultation where a successful (below target) blood pressure reading is taken relative to other consultations. So, the volume of primary care consultations (x_j) is weighted upwards if any feature successful blood pressure management, with the measure of primary care output being formulated as:

$$\Delta Y_{jt}^{pc} = \frac{x_{jt} \left[0.3 * \sum_{m=1}^{3} P_{mt} S_{mt} \right]}{x_{jt-1} \left[0.3 * \sum_{m=1}^{3} P_{mt-1} S_{mt-1} \right]} \tag{2.3}$$

Where m indexes the three conditions $m = 1 \dots 3$, P_m is the prevalence rate for condition m and S_m is the QOF success rate for condition m.

Table 1 provides an overview of the current information used to account for the quality of NHS output, and shows which settings are covered.

Table 1: Quality adjustments for cost-weighted output

Setting	NON-ELECTIVE	ELECTIVE / DAY CASE
INPATIENT	30-day survival (by HRGs) Remaining life expectancy (by HRGs)	30-day survival (by HRGs) Remaining life expectancy (by HRGs) 80th percentile of waiting times (by HRGs)
OUTPATIENT		80th percentile of waiting times[1]
PRIMARY CARE		QOF Blood Pressure indicators: Chronic Heart Disease Stroke Hypertension

Key: HRG: Healthcare Resource Groups; QOF: quality and outcomes framework.
Sources: [12, 13]

2.3 An alternative source of QALYs

Whilst there are no routine national datasets of QALYs for NHS patients, there is an alternative source that should be considered for its potential relevance, namely data from a project that sought to develop methods for estimating the NICE cost-effectiveness threshold [14]. The project aimed to estimate the relationship between changes in overall NHS expenditure and changes in mortality and to translate these estimates into broader effects in terms of QALYs. To support this aim, one of the objectives of the work was to estimate "the quality of life (QoL) associated with additional years of life and the direct impact of health services on QoL" In theory, estimates from this project could replace some of the arguments in equation 2.2 above.

There were two key elements to the QALY estimates constructed as part of the project. The first element consists of estimated effects of changes in NHS expenditure on mortality, described in terms of years of life gained (*YLG*). These gains reflect how changes in health expenditure impact on mortality and take "into account the 'counterfactual' deaths that would have occurred if the population [in a particular expenditure category] … faced the same mortality risks as the general population" [14]. The second element adjusts the *YLG* estimates to reflect how QoL differs by age and gender. There are two forms of QoL adjustment, differing according to the source of the data.

The first form, Y_{dg}^{Q1}, uses data from the Health Outcome Data Repository (HODaR) [15] which "provides over 30,000 observations of EQ-5D[2] measures of quality of life by ICD10[3] code and the age and gender of the patients in the sample" supplemented with information from the Medical Expenditure Panel Survey (MEPS). For the project, "these data provided a means of estimating the quality of life associated with each ICD code at the average age of respondents in the pooled sample" [14]. The data are available online.[4]

This form of the adjustment can be written:

$$Y_{dg}^{Q1} = YLG_{dg}h_d^1 \tag{2.4}$$

---

[1] See Dawson 2005, CHE RP6, p96 (section 4.10.5; Table 4.7)
[2] EuroQol five dimensions questionnaire https://euroqol.org/
[3] International Statistical Classification of Diseases and Related Health Problems 10th Revision
[4] https://www.york.ac.uk/media/che/documents/Appendix%20A%20(Displacement%20by%20ICD%20code).xlsx

Where Y_{dg}^{Q1} indicates the first form of the QALY adjustment, *d* denotes diagnostic (ICD10) groups and *g* 5-year age and gender groups. YLG_{dg} indicates the years of life gained, by diagnostic and age/gender band. In this formulation, the estimates of *YLG* are weighted according to the quality of life h_d^1 HODaR/MEPS survey responses by diagnostic group (but not by age and gender, because the pooled sample was too small).

The second form, Y_{dg}^{Q2}, applies "QoL 'norms' for the general population by age and gender based on an analysis of data from the Health Survey for England (HSE)." The resulting QALY formulation can be written as:

$$Y_{dg}^{Q2} = YLG_{dg}h_g^2 \tag{2.5}$$

The estimates of *YLG* are weighted using the HSE health-related quality of life h_g^2 norms for the general population by age and gender (but not by diagnosis).

It may be that the data used to construct these two calculations of QALYs could be used in the quality adjustment measure for calculation productivity growth. We consider the estimation of mortality effects and of quality of life effects in turn.

Mortality effects

We could substitute the YLG_{dg} estimates for the life expectancy estimates L_j (which also take account of the age/gender composition so, comparably, can be written L_{gj}).

There are two potential advantages to using YLG_{dg} estimates. First, as YLG_{dg} estimates are available by ICD10 codes, these estimates could be applied to match the diagnostic composition of patients within each HRG, to arrive at estimates of YLG_j. This would allow changes in the diagnostic composition of patients in each HRG to be captured, rather than changes in just the age composition, as currently.

Second, the YLG_{dg} estimates attempt to capture the "effects on mortality to life years taking into account the 'counterfactual' deaths that would have occurred if the population … faced the same mortality risks as the general population". The advantage is critical when making comparisons across disease areas, because comparisons must be based on the QALYs gained as a consequence of treatment, relative to what would have been experienced in the absence of treatment.

Calculation of QALY gains is less critical, though, when measuring changes over time in mortality or life expectancy, as required in measuring output growth, when it is necessary only to assess changes in the with-treatment health profile over time, under the assumption that the counterfactual – what happens in the absence of treatment – is constant over time for each patient group [3, 7].

The disadvantage of using YLG_{dg} is that these data are not strictly comparable to our life expectancy estimates L_j. The mortality effects YLG_{dg} are estimates of the impact on mortality of a £1 change in NHS expenditure, built up from estimates of (i) the deaths averted per pound spent and translated into (ii) life years gained according to the age at which a disease-specific death typically occurs. The first component is not required for the measurement of output growth. The second element is based on the ONS life tables [10], just as for L_j in equation 2.2, the difference being that the former are applied by disease-group rather than HRG. Thus, with respect to the life expectancy effects, the NICE cost-effectiveness threshold project and the life expectancy formulation in equation 2.2 adopt a broadly equivalent approach and the same data.

Quality of life effects

We could substitute either h_d^1 or h_g^2 for k_j in equation 2.2. Note that none of these arguments include a time dimension, with values not varying over time. This means that the choice of which argument to use boils down to a preference about the quality of the QoL measures.

In our current practice, for only a handful of HRGs are actual QoL data available, these being the four conditions for which Patient Reported Outcome Measures are collected. For all other patients treated in hospital, we assume that the ratio of average health status before and after treatment, amounts to 0.8 for elective patients and 0.4 for non-elective patients and that these ratios remain constant over time. On the face of it, relaxing these assumptions using either h_d^1 or h_g^2 would seem advantageous. But this depends on the quality of the underlying data.

The Health Outcome Data Repository (HODaR) provides over 30,000 observations of EQ-5D measures of quality of life by ICD code. The data were derived from patients "treated at Cardiff and Vale NHS hospital from 2002 to 2004. Inpatients were surveyed 6 weeks post-discharge whilst outpatients are handed a survey package when they attend" (ref p74, threshold project report).

On the face of it these data could substituted for k_j by applying the h_d^1 estimates to each patient according to their primary diagnosis, and aggregating these to HRGs to derive estimates at HRG level, h_{dj}^1. In our original work [3] we considered this possibility, and examined the quality of the HODaR data. We found that the data do not constitute before and after measures of health status, with multiple observations only available for a fairly small proportion of patients. Moreover, for those with multiple observations, the time intervals varied considerably and the ICD10 codes often changed from one survey to the next, suggesting that they are receiving treatment for different underlying conditions. We concluded that the data were not fit for purpose because it was not clear what the HODaR data were actually capturing:

We have analysed the HODaR set of observational data to ascertain whether the information can be utilised in the construction of outcome weights for a productivity index. We have concluded that the HODaR data are unsuitable for this purpose. The surveys have not been administered with the express intention of collecting before and after information. Although multiple surveys exist for a subset of patients, it is unlikely that many of these constitute before-and-after measurements. (Page 64, Technical appendices to [3])

For the NICE cost-effectiveness threshold project, the HoDAR/MEPS data were interpreted as providing a measure of QoL associated with the disease under current care arrangements but not as providing a measure of the *changes* in QoL as a consequence of treatment. The use of h_d^1 in the NHS output measure, therefore, does not seem worth pursuing.

Estimates of h_g^2 are derived from the Health Survey for England (HSE), which is designed to be representative of the English population. These estimates were derived after pooling data from six surveys (1996, 2003-2006 and 2008).

The estimates could be applied to reflect changes in the age/gender composition of patients in each HRG each year, to arrive at an estimate of h_{gj}^2. This would involve substituting k_j for h_{gj}^2 in equation 2.2. This would be an improvement on our current practice of using estimates for all elective and all non-elective patients. But the underlying QoL estimates would remain time-invariant, deriving from the pooled HSE surveys. For it to be worth moving to this form of quality adjustment, at the very least the estimates of h_g^2 would need to be updated, given that the most recent survey data on

which they are based date from 2008. Moreover, as with HoDAR/MEPS data, the estimates of h_g^2 do not provide a measure of the *changes* in QoL as a consequence of treatment.

In summary, therefore, the data used for the NICE cost-effectiveness threshold project offer insufficient improvement to our current quality adjustment approach. A better avenue might be to seek to supplement our quality adjustment with indicators that measure different elements of quality not captured directly as QALYs. We explore this in the remainder of the report.

3. Review of existing criteria for indicator selection

In the absence of off-the-shelf and regularly updated measures of the change in QALYs associated with NHS treatment, and having ruled out the NICE cost-effectiveness threshold project as a viable alternative, we then considered whether and how other routine indicators of the quality of care could be incorporated into the measure of NHS output growth. This involved establishing a set of criteria that potential indicators ought to satisfy if they are to be considered for inclusion. We set about establishing these criteria in two stages.

First, we searched for criteria that had been developed by national bodies responsible for collating and publishing measures of patient safety and quality. These criteria are not about the type of quality (i.e. the content of the indicators) but rather relate to the properties of indicators that make them suitable for measuring quality. We undertook a detailed review of three of these criteria sets:

- AHRQ 2008 (Agency for Healthcare Research and Quality (US))
- NCHOD 2005 (National Centre for Health Outcomes Development)
- HSCIC 2014 (Health and Social Care Information Centre)

Second we presented our findings from this review at a workshop attended by UK experts in healthcare quality and/or in the measurement of healthcare output and productivity indices. We asked participants to suggest the criteria that indicators should satisfy if they are to be included in an NHS output index.

3.1 AHRQ criteria used to evaluate potential quality indicators

The Agency for Healthcare Research and Quality (AHRQ) is a US government agency responsible for improving the safety and quality of America's healthcare. It has developed a range of evidence-based quality indicators to assess performance, and to identify variations in care quality. AHRQ quality indicators are updated annually and are frequently used in research projects as well as for comparative reporting and performance assessment within the US healthcare system [16]. In 1998, researchers at University of California, San Francisco (UCSF) and the Stanford University Evidence-Based Practice Center (EPC) were commissioned to review and revise the AHRQ's existing set of indicators [17]. Based on a literature review, and informed by interviews with a wide range of stakeholders (including academics), the researchers produced a set of criteria, see (Table 2) for evaluating potentially viable indicators [16].

Table 2: AHRQ criteria

Face validity	An adequate quality indicator must have sound clinical or empirical rationale for its use. It should measure an important aspect of quality that is subject to provider or healthcare system control.
Precision	An adequate quality indicator should have relatively large variation among providers or areas that is not due to random variation or patient characteristics. This criterion measures the impact of chance on apparent provider or community health system performance.
Minimum bias	The indicator should not be affected by systematic differences in patient case mix, including disease severity and comorbidity. In cases where such systematic differences exist, an adequate risk-adjustment system should be possible using available data.
Construct validity	The indicator should be related to other indicators or measures intended to measure the same or related aspects of quality. For example, improved performance on measures of inpatient care (such as adherence to specific evidence-based treatment guidelines) ought to be associated with reduced patient complication rates.
Fosters real quality improvement	The indicator should be robust to possible provider manipulation of the system. In other words, the indicator should be insulated from perverse incentives for providers to improve their reported performance by avoiding difficult or complex cases, or by other responses that do not improve quality of care.
Application	The indicator should have been used in the past or have high potential for working well with other indicators. Sometimes looking at groups of indicators together is likely to provide a more complete picture of quality.

Source: AHRQ. *Refinement of the HCUP Quality Indicators*, p. 30.

3.2 NCHOD Criteria (matrix) for evaluating the quality of indicators

In 2005, the National Centre for Health Outcomes Development (NCHOD) published a new set of criteria ('matrix') for evaluating the quality of clinical and health indicators [18]. Reviewing 18 independent sources on criteria and methods, NCHOD organised the resulting indicators into four groups: scientific criteria; policy criteria; methodological criteria; and statistical criteria. A summary of these criteria is presented in Table 3**.**

These four sets of criteria are further subdivided into three phases of the indicator's life cycle:

- Development (scientific criteria; policy criteria)
- Measurement (methodological criteria)
- Interpretation (statistical criteria)

These phases are progressive, i.e. an indicator must satisfy the 'development' phase before progressing to assessment at the higher levels. The 'measurement' phase must be satisfied before criteria in the 'interpretation' phase are applied. Indicators are then assessed against each quality criteria using a 5-star rating system (* =v. poor; ***** =v. good).

Table 3: NCHOD Criteria

Scientific criteria (SC)		Code
Explicit definition	Is the indicator explicitly defined by appropriate statistical units of measurement and clinical terminology?	SC1
Indicator validity	Will the indicator measure the phenomenon it purports to measure i.e. does it makes sense both logically and clinically?	SC2
Scientific soundness	How scientific is the evidence / selection process (systematic / non-systematic) to support the validity of the indicator?	SC3
Policy criteria		
Policy-relevance	Does the phenomenon under measurement represent significant public interest, disease burden or cost?	P1
Actionability	Can the factors which influence the phenomenon be positively influenced to induce a future health / cost benefit?	P2
Perverse incentives	Will the measurement process encourage undesired behaviours by those under measurement?	P3
Methodological criteria (M)		
Explicit methodology	Are measurement tools / procedures explicitly defined, understood and monitored?	M1
Attributability	Are the factors which influence (+/-ve) the phenomenon likely to be identified e.g. patient risk factors, practitioner procedure etc?	M2
Timeliness	What is the average time (months) between measurement and results?	M3
Frequency	What is the average time (months) between reporting of results?	M4
Sensitivity to change	Do the measurement tools and timing of results allow changes to be observed over time?	M5
Confounding	What is the risk that variations between organisations and changes over time may be influenced by confounding factors?	M6
Acceptability	What % stakeholders accept the process of measurement and the reasons for it?	M7
Measurability	Is the measurement process possible within the available budget and resources?	M8
Cost-effectiveness	Does the likely output represent a cost-effective use of budget/resources?	M9
Statistical criteria (SP)		
Specificity	Does the measurement appropriately capture the level of detail required e.g. sub-group analyses, accurate diagnosis?	SP1
Comparability	Is the measure comparable between relevant sub-groups e.g. are age/sex/geography-specific data standardised and consistent?	SP2
Representativeness	Are sample sizes representative across all required sub-groups	SP3
Data quality	% of the information missing from the records?	SP4
Data reliability	% agreement (kappa coefficient) between measured records and those collected by an independent source?	SP5
Uncertainty	Have appropriate techniques been selected to demonstrate the effects of variation, dispersion and uncertainty (Shewhert, funnel plots etc.)?	SP6
Interpretability	Can understandable, meaningful and communicable conclusions be drawn from the results?	SP7

Source: NCHOD's matrix (the codes are our own abbreviations).

3.3 HSCIC (Indicator Assurance Service)

The Health and Social Care Information Centre[5] (HSCIC) is the national provider of information, data and IT systems for health and social care In England. HSCIC has an assurance process for determining the suitability of quality indicators for use in the National Library of Quality Assured Indicators.[6] The HSCIC criteria are [19] [7]

- **Clarity:** Is it clear what the indicator will measure?
- **Rationale:** What are the reasons and evidence for measuring this?
- **Data:** Is the data in the measure fit (enough) to support the purpose?
- **Construction:** Will the methods used support the stated purpose? Is it clear what methods are used and how they have been tested and justified?
- **Interpretation:** Is the presentation of the indicator suitable and are all potential users able to interpret the values? Can the indicator be used for quality improvements?
- **Risks:** Are any limitations, risks or perverse incentives associated with the indicator explicitly stated?

Each of these headline criteria is accompanied by further explanatory statements. These are reproduced in **Table 4**.

Table 4: HSCIC List of Criteria (2014)

1. Clarity: Is it clear what the indicator will measure?	
a)	A unique name for the measure which is sufficiently descriptive to convey meaning when referenced or quoted without supporting meta-data and differentiates it from, or specifically associates it with, other indicators.
b)	A clear and unambiguous description of the measure, which is expressed both in plain English and the relevant clinical and/or statistical terminology of the particular subject in question, and which is suitable for a diverse audience.
c)	A clear statement about the measurement units, and reasons why that unit has been chosen as relevant.
d)	A clear statement about the scope of the indicator, which will typically include aspects such as detailed patient, population, disease group, geographical and geographical granularity coverage.
e)	All other major inclusions and exclusions should be stated in the indicator definition.
2. Rationale: Are the reasons and evidence for measuring this clear?	
a)	The sponsor for the measure should be clearly stated.
b)	A clear statement about the purpose of the measure.
c)	A clear identified gap or need for the indicator.
d)	Justification as to why this is a sufficiently important question/service that requires measurement.
e)	A clear statement about the evidence base for the measure such as clinical evidence or professional consensus, and if relevant it should be acceptable to those whose behaviour and practices this may be applied.
f)	A clear statement of the policy objective and/or critical business question that the measure is seeking to capture. The rationale must be clearly set out, be plausible, and capable of being understood by a diverse audience including the public.
g)	If the indicator fits into a framework, the rationale for the framework as a whole and an outline of how the indicator is included.
h)	Previous decision-making documents are included for reference.
3. Data: Is the data in the measure fit to support the purpose?	
a)	The source of the data is clearly identified with justification, including the extent of any intermediate processing steps which might predispose the data to errors or bias. How data will be extracted or collected is included, with justification if required.

---

[5] The Health and Social Care Information Centre changed name to NHS Digital in April 2016.
[6] http://content.digital.nhs.uk/article/5173/Appraisal-end-to-end-process
[7] http://content.digital.nhs.uk/article/5171/Quality-Criteria

b)	Whether the indicator data source is re-using a collection/extraction or is primarily being collected / extracted for the indicator is discussed.
c)	Alternative data sources have been considered with justification as to why they were not used.
d)	Data availability is discussed, including the form in which it is available, who has access to the data and evidence that it is available with sufficient frequency and timeliness to enable desired improvement actions to be visible. The availability of data long-term has been considered.
e)	The data used is robust enough to support the measure and its derivations. The quality of the data is above the threshold of acceptability, and this threshold is explicitly defined in the method, and accepted by all stakeholders. The effect of data quality issues upon the measure are explicitly known and declared.
f)	An explicit definition of any exclusions from the scope, (which might include specific instances, or be based on calculated or derived rules) along with justification as to why these have been excluded.

4. Construction: Will the methods used support the stated purpose? Is it clear what methods are used and how they have been tested and justified?

a)	The measure construction, and/or relevant derivations from it are explicitly defined and justified, to the extent that it is possible to reconstruct the measure and/or derivations using the same base data.
b)	The construction of the indicator is fit for purpose and supports the stated rationale.
c)	The element of chance has been appropriately considered in the design of the measure, and in any associated derivations or statistical models.
d)	Indicator is sensitive to changes in true events.
e)	An assessment has been made of the relevance and significance of case-mix, risk, age and sex adjustments in the context of the business question / improvement objective, or any other adjustments relevant to the indicator. An explanation as to what extent these have been carried out and any testing used to inform choice of standardisation method used (if relevant) should be summarised.
f)	The use of confidence intervals or control limits has been stated, with the relevant methodology and justification.

5. Presentation and Interpretation: Is the presentation of the indicator suitable and are all potential users able to interpret the values? Can the indicator be used for quality improvements?

a)	Consideration of whether any contextual information is required to accurately interpret the indicator. Construction of appropriate contextual information is presented.
b)	An explanation is provided as to whether targets or target ranges will be used with supporting evidence of how these are derived. Where targets are not used, how direction of travel should be interpreted by the user is provided.
c)	The indicator is capable of detecting variability that is important enough to warrant further investigation.
d)	Clear statement regarding how the indicator should be used and how it can be used for comparison. Clear explanation of when the indicator cannot be used, with justification.
e)	A list of caveats to be presented with the indicator has been included. A thorough investigation into limitations has been carried out and has been addressed as successfully as possible.
f)	Any biases resulting from scope, sample size or data collection/extraction factors have been clearly identified.
g)	Consideration has been given to the forms of presentation of the indicator for the intended stakeholder audience. These are appropriate and have been tested or verified in some way.
h)	Any common industry standard conventions for presentation have been adopted e.g. standard error bars, labelling, scale, limitations, exclusions etc.
i)	To what extent action can be taken to improve a 'bad' position suggested by an adverse indication is clearly stated, and what steps can be taken to improve the measurement. Providers and commissioners are able to improve the results of the measurement

6. Risks: Are any limitations, risks or perverse incentives associated with the indicator explicitly stated?

a)	A purpose and description of any similar existing indicators are presented alongside justification as to why an additional indicator is needed. Differences in purpose and construct are clear and appropriate.
b)	Methodology is consistent with other existing indicators or indicators within the same set, or justification is provided as to why this is not appropriate.
c)	Consideration as to whether results of the measurement would contradict other existing indicators and any resulting impacts of this.
d)	If the measure, or the process of measurement, introduces undesired behaviours by those being measured, these are clearly stated. If the extent of this is known or predictable, it does not invalidate utility of the indicator.

e)	To what extent the indicator is susceptible to the risk of 'gaming' is clearly stated, outlining whether the measure is capable of being manipulated in some way to influence the outcome without the intended improvement actions taking place.
f)	Issues around disclosure control have been considered.

Source: HSCIC publication 'Criteria and considerations used to determine a quality indicator', Indicator Assurance Service, (2014).

3.4 Expert Workshop

In June 2016, the research team held a one day workshop with key experts in the area of health system productivity measurement including representatives from the Department of Health, the Office for National Statistics, The Health Foundation, the Nuffield Trust, Health Education England and the Care Quality Commission.

The primary objective of the workshop was to seek expert views on the criteria that a quality indicator ought to satisfy in order to be incorporated in the measure of NHS output and productivity. Workshop participants were tasked to come up with their own list of key criteria that an indicator of quality should have, and to discuss these as a group exercise. Three groups were formed, each attended by a member of the research team. In particular, participants were posed the following question:

"Which CRITERIA should an indicator measuring the quality of the health care system satisfy to make it suitable for use in an output index?"

After this exercise, each group was asked to list the criteria/properties they had identified and to provide a short explanation as to why they were chosen and their relevance. These were related to all workshop participants by a spokesperson for each group and moderated by a member of the research team. Summaries of discussions were taken by each member of the research team overseeing the three group discussions. The full list of criteria proposed by workshop participants is reported in Box 1.

Box 1 - Criteria suggested by workshop participants

1. Clarity
2. Not ambiguous
3. Uni-directional
4. Appropriately defined area of activity
5. Value to patient / Relevant for the patient
6. Value to clinicians
7. Inform patient safety/improvement
8. Coverage/comprehensiveness at national level
9. Granularity
10. Adjustable for different types of populations
11. Adjustable to specific use and sector
12. Consistency over time
13. Timeliness
14. Ability to capture meaningful change
15. Specificity
16. Sensitivity

The two main take-home messages of the workshop were that all criteria discussed at the workshop should be considered as equally important and that at least one indicator of the quality of healthcare delivered in each NHS settings (starting from those settings for which no quality dimension is currently captured) should be included in the measure of NHS output and productivity, provided that it is available and that it satisfies the criteria set. A warning was also provided to be careful when considering 'satisfaction measures' of health care services as these are usually subjective measures.

4. Criteria for quality indicators in output measures

Following the workshop, we finalised our set of criteria, drawing together commonalities across the reviewed criteria sets and moderated by the opinion expressed by and discussions held by workshop participants. This generated a set of seven criteria that indicators should satisfy if they are to be included in a measure of output growth.

After reviewing the three published criteria sets and sharing these at the workshop, we derived a set of criteria that measures of patient safety and quality ought to satisfy if they are to be considered for inclusion in a measure of NHS output. Our list of seven criteria is:

- Clarity
- Added value
- Benefit / measure of value
- Attribution / granularity
- Consistency over time
- Measurability / timeliness
- Validity

Table 5 shows how the published criteria sets reviewed in Section 4 and those suggested at the workshop map to the seven criteria. Each of these is discussed below.

Table 5: List of seven criteria, and cross mapping with existing quality indicator criteria sets

Criterion	Wording	AHRQ 2008	NCHOD 2005	HSIC 2014	Workshop criteria
Clarity	An indicator should be specific, stating clearly what is being measured, how it is measured and its interpretation.		SC1 M1 SP1, SP7	1. Clarity 4. Construction (a) 5. Interpretation	Clarity Appropriately defined area of activity Not ambiguous Unidirectional
Added value	The indicator should capture elements of quality not captured elsewhere.	Construct validity Application		2. Rationale (e, g) 6. Risks (a, c)	
Benefit / Measure of value	The indicator should capture aspects of quality proven to be valued by users of health services.	Face validity	P1 M7	2. Rationale (a, e, f)	Value / relevant to patient Value to clinicians Inform patient safety/improvement
Attribution / Granularity	The indicator should measure aspects of quality that can be attributed to healthcare settings and that are subject to health system control.	Face validity Fosters real quality improvement	P2 M2, M8, M9	4. Construction (b, c, d) 5. Interpretation (i)	Coverage and comprehensiveness at national level Granularity - adjustable to different target populations, settings and to specific use
Consistency over time	The indicator should measure quality aspects consistently over time, at least annually, and in a timely manner.		M4, M5, M6	3.Data (d)	Consistency over time
Measurability / Timeliness	The quality aspect should be measurable / quantifiable.	Precision	M3	3.Data (e)	Timeliness
Validity	The indicator measures what it intends to measure, i.e. it is a discriminates between good and bad quality, is not subject to large variation due to random changes in small numbers of events, and minimises potential bias.	Precision Minimum bias Construct validity	SC2, SC3 P3 SP2, SP3, SP4, SP5, SP6	4. Construction (a – reproducibility; (e, f – risk adjustment; uncertainty) 5. Interpretation (d, sensitivity; e, caveats) 6. Risks	Specificity Sensitivity Ability to capture meaningful change

4.1 Clarity

> **Clarity**
> An indicator should be specific, stating clearly what is being measured, how it is measured and its interpretation.

The need for an indicator to be clearly defined is a logical necessity, and two published criteria sets include this property. The HSCIC set focuses on the clarity of what is measured, covering the indicator's name, description, measurement units, and scope (1), its construction (4a), and interpretation (5). The NCHOD matrix requires indicators to have an explicit definition, using appropriate statistical units and clinical terminology (SC1), an explicit methodology (M1), the appropriate unit of analysis (SP1) and a clear interpretation (SP7). The AHRQ set does not specify the need for clarity.

Workshop participants stressed that indicators should clearly state the purpose for which the indicators are to be used. Further, workshop participants stated that indicators should satisfy specific statistical properties, such as being unidirectional and able to deal with volatility of the data.

4.2. Added value

> **Added value**
> The indicator should capture elements of quality not captured elsewhere.

The NCHOD criteria do not include the notion of 'added value'. HSCIC specifies that an indicator should fill an unmet need or 'gap' (2e), and its role within the performance framework should be justified (2g; 6a, c). AHRQ addresses 'added value' under the heading 'application'. This captures the need for a single indicator to 'work well' with other indicators in a performance framework, so that together they provide a fuller picture of quality.

The idea of 'added value' is more important for purposes of output measurement than it might be for construction and reporting of quality indicators for assessing performance. This is because the output measure should be designed to capture comprehensively the bundle of each product's characteristics but needs to minimise the risk of 'double-counting' these characteristics wherever possible.

In order to evaluate the risk of double-counting, the entire set of candidate quality measures needs to be jointly assessed, rather than considered in isolation. If two partially overlapping indicators are both to be incorporated in the output measure, there needs to be careful thought about what relative value to assign to each of these when they are combined so that their distinct contributions are accurately captured.

4.3 Benefit / Measure of value

Benefit / Measure of value
The indicator should capture aspects of quality proven to be valued by society.

All three published criteria sets include this concept. The HSCIC set specifies that the indicator's rationale should refer to the sponsor (2a), providers (2e) and policy makers (2f). The NCHOD matrix includes policy relevance ('public interest', P1) and stakeholder acceptability (M7). Relevant stakeholders include users (patients), commissioners, providers and policy makers – the public and/or tax payers could also be included. The AHRQ criterion of 'Face validity' includes the need for a sound clinical or empirical rationale.

Workshop participants agreed on the importance of this criterion, stressing that indicators should capture aspects of quality that are valued by and relevant to patients (or by clinicians and other patient representatives, in recognition of the agency relationship in healthcare).

Here again, there is divergence between the published criteria and those required for national accounting purposes. The national accounting measure of health output is supposed to capture the social value of what the health service produces. This means that output should reflect the value that society places on the activities of the health system. Hence, the scope of this criterion is perhaps more focused than that expressed in other criteria sets.

4.4 Attribution / granularity

Attribution / granularity
The indicator should measure aspects of quality that can be attributed to healthcare settings and that are subject to health system control.

An indicator is unsuitable as a measure of the quality of health care output if it represents a phenomenon outside the control or influence of the health system. All three criteria sets incorporated this idea explicitly. AHRQ classifies this concept under its 'Face validity' criterion, specifying that an indicator should be subject to provider or healthcare system control. Also in the AHRQ set, 'Fosters real quality improvement' covers genuine engagement and the need to minimise incentives for gaming behaviours.

In the 'construction' section of its criteria, HSCIC requires indicators to be fit for purpose (4b), take account of chance (4c) and be sensitive to changes in true events (4d). Another relevant HSCIC criterion is the ability of healthcare providers to improve their performance (5i); HSCIC also suggests that there should be some guidance alongside the indicator that specifies the steps that commissioners and providers need to take.

The NCHOD matrix uses the term 'actionability' to describe provider capacity to positively influence the indicator and so 'induce a future health/ cost benefit' (P2). Attributability is one of the matrix's methodological criteria (M2) as are measurability (M8) and cost-effectiveness (M9). These last two concepts underscore the need for provider engagement in designing the indicator, to ensure that the measurement process is feasible and affordable.

Workshop participants also stressed that, for this criterion, there should be clarity about what is attributable to the actions of different parts of the health system. For example, the health system as a whole might contribute to improved life expectancy, but these improvements reflect the combined efforts of different parts of the health system, not just hospitals or primary care acting in isolation. Participants suggested that indicators of healthcare quality should be adjustable to specific uses and

settings and for different target populations. For instance, indicators should be capable of disaggregation to different units of analysis, eg national versus Trust level productivity measurement. These sentiments imply the need for a common set of process and outcome indicators, to be collected in relevant healthcare settings.

Also, it was suggested that attribution is much more problematic when considering outcome measures than process measures of the quality of healthcare. This is because health outcomes are often determined by wider social factors and not only (or, often, even primarily) by healthcare utilisation. Ideally the contributions of these wider determinants would be captured in other parts of the national accounts. For example, the beneficial influences on health of improvements in the housing stock ought to be captured in the measure of housing's contribution to GDP.

In summary, for national accounting purposes, it is important that the output measure captures the valuable characteristics and consequences of healthcare activities, not of other sectors of the economy.

4.5 Consistency over time

> **Consistency over time**
> The indicator should measure quality aspects consistently over time.

Usually quality measures are constructed so as to monitor progress over time, and this is more easily accomplished if the indicators are measured in a consistent fashion from one period to the next. Similarly in national accounting, the objective is not merely to measure the level of output at any particular point in time but to measure changes in output over time.

The ARHQ criteria do not explicitly include the need for an indicator to measure quality consistently over time. In contrast, the HSCIC criteria specify the importance of both the frequency and timeliness of data available to calculate the indicator (3d). The NCHOD matrix differentiates three aspects of timeliness: the reporting frequency (M4), whether the indicator captures longitudinal changes in quality (M5) and whether other time-varying factors can be recognised and adjusted for (M6).

Workshop participants also said it was important that indicators should be measured on a timely and consistent basis.

The requirement for consistency over time is less demanding for national accounting purposes than it is for reporting of quality measures, where consistent definitions are required in order to construct a time series. But in national accounting a long time series is usually constructed as a 'chain index'. The links in the chain only require that outputs are measured in a consistent fashion across two successive periods. If there are definitional changes, the chain index requires that outputs are measured using both old and new definitions in the period that the change occurs, so that a link to both the past and the future can be constructed.

Timeliness of a quality indicator is also very important for the purposes of NHS output and productivity measurement, as series updates are usually conducted on an annual basis. Therefore, it is vital that quality indicators are available within the same timeframe.

4.6 Measurability / Timeliness

Measurability / Timeliness
The quality aspect should be measurable / quantifiable, at least annually, and in a timely manner.

Whilst accepting the adage that what matters and what can be measured are not necessarily the same, any quality indicator has to be based on the measurable. All three criteria sets incorporate this requirement, with HSCIC summarising the concept as the need for data to be available and "robust enough to support the measure and its derivations" (3e). This also applies for national accounting purposes.

Workshop participants also suggested that indicators of healthcare quality should be comprehensive at the national level and available on a timely basis.

4.7 Validity

Validity
The indicator measures what it intends to measure.

For the purposes of constructing a national measure of healthcare output, the quality measure needs to be valid. More fully, it needs to be able to discriminate between good and bad quality, is not subject to large variation due to random changes in small numbers of events, and minimises potential bias. All three of the published criteria sets reviewed in Section 4 stated that indicators needed to satisfy one or other conceptualisation of validity.

Workshop participants also stressed the importance of both specificity and sensitivity and the ability of any indicator to capture 'meaningful change'.

If the output measure is to be employed sub-nationally, perhaps to understand and manage performance of healthcare organisations, then additional considerations of sensitivity and specificity might need to be considered. Sensitivity concerns the ability of the indicator to detect true positive values and specificity its ability to detect true negative values. Typically there is a trade-off between sensitivity and specificity (i.e. improved sensitivity might entail lower specificity).

5. Do published indicators satisfy the criteria?

5.1 Introduction

The final objective of this work was to assess the extent to which published indicators of the quality of healthcare satisfy the seven criteria, with the aim of finding indicators for potential use in the NHS output measure. We focussed on the NHS Outcome Framework (OF) indicators and the NHS Safety Thermometer indicators, which are summarised in the section 5.2. A full summary of each of the NHS Outcomes Framework and the NHS Thermometer indicators is provided in Appendix B.

The review of the indicators against the seven criteria was performed as an individual task by members of the research team, and also by two experts, one from the DH and one from ONS.

Each reviewer was provided with an excel spreadsheet set up as a matrix listing all NHS Outcome Framework (OF) indicators / NHS Safety Thermometer indicators against the seven criteria on which to record. Reviewers were asked to indicate whether or not they thought the indicator satisfied each criterion, whether an indicator should be included in the NHS output index and to select an appropriate NHS setting (Main and secondary) to which they thought the indicator related to. A blank box was also added for reviewers to record any comments they might have.

The excel spreadsheet also contained an instructions sheet and a Criteria & definition sheet. To ease the reviewing process, each reviewer was sent a document including a brief summary of the criteria selection process, a summary of the published criteria set and a summary of the seven criteria identified by the research team. Reviewers were also sent a summary of all NHS OF indicators and NHS Safety Thermometer indicators, providing information about the definition, purpose and data availability for each indicator. Reviewers were asked to complete the exercise as an individual task. Completed spreadsheets were collected and collated, and summary tables produced, these being summarised in section 5.3.

5.2 Published indicators

NHS Outcomes Framework Indicators

The NHS Outcomes Framework (OF) Indicators were developed in December 2010 with the aim of providing national accountability for the outcomes delivered by the NHS. Its objective is threefold: to be a driver of transparency, of improvement in the quality of care and of measurement of the outcomes produced by the NHS. The focus is on improving health and reducing health inequalities. The NHS OF delineates national outcomes goals, which are then used to monitor the progress of the NHS in England. [8]

NHS OF indicators are grouped in five domains, which list the "high-level national outcomes that the NHS should be aiming to improve".[9] Each domain has a number of overarching indicators and a number of improvement areas.

---

[8] https://beta.digital.nhs.uk/publications/ci-hub/nhs-outcomes-framework (last accessed 6th March 2018)
[9] https://beta.digital.nhs.uk/publications/ci-hub/nhs-outcomes-framework (last accessed 6th March 2018)

The five domains are:

Domain 1 - Preventing people from dying prematurely
This domain captures how successful the NHS is in reducing the number of avoidable deaths.

Domain 2 - Enhancing quality of life for people with long-term conditions
This domain captures how successfully the NHS is supporting people with long-term conditions to live as normal a life as possible.

Domain 3 - Helping people to recover from episodes of ill health or following injury
This domain captures how people recover from ill health or injury and wherever possible how these can be prevented.

Domain 4 - Ensuring that people have a positive experience of care
This domain looks at the importance of providing a positive experience of care for patients, service users and carers.

Domain 5 - Treating and caring for people in a safe environment and protecting them from avoidable harm
This domain explores patient safety and its importance in terms of quality of care to deliver better health outcomes.

All indicators are calculated separately and where possible by local authority (both lower tier and upper tier), region, age, gender, sexual orientation, ethnicity, religion, deprivation decile and condition.

NHS Thermometer Indicators
The NHS Safety Thermometer is the measurement tool for a programme of work to improve patient safety. Indicators were collected and reported by NHS Digital from 2012/13 until April 2017. Since then, the NHS South, Central and West Commissioning Support Unit (SCW) has been managing the collection and publication of NHS Thermometer indicators on behalf of NHS Improvement.

The NHS Thermometer indicators are used to record patient harms at the frontline, and to provide immediate information and analyses for frontline teams to monitor their performance in delivering harm free care, thereby leading to the reduction and, hopefully, elimination of, harm. Patients are assessed in the setting in which they receive care.

The first indicators were introduced in 2012/13 and covered four areas, which were specifically relevant to older people who, experiencing more healthcare intervention, were more at risk of multiple harms. The four areas are:

- Pressure ulcers
- Falls
- Urinary tract infections (UTIs) in patients with a catheter
- New venous thromboembolisms (VTEs)

Taken together, these indicators are also known as the 'classic' NHS thermometer. These four harms were prioritised for attention by the Department of Health's Quality, Innovation, Productivity and Prevention (QIPP) Safe Care programme because they are common, and because there is a clinical consensus that they are largely preventable through appropriate patient care. The concept of a

composite measure, Harm Free Care, was designed to bring focus to the patient's overall experience.

Subsequently patient safety indicators were developed for other areas of health care, namely:

- Maternity care[10]
- Medication[11]
- Mental Health services[12]
- Children and Young Peoples' services

The box below reports the types of potential harms about which indicators are collected under these four broad areas.

Maternity Care	Medication
· Maternal Infection · Perineal Trauma · Post-Partum Haemorrhage · Term babies Apgar score · Term baby treatment [no data presented] · Women's perception of safety	· Medication Reconciliation (MR) · Allergy status · Medication omission · Omissions of Critical Medication · Identifying harm from high risk medicines
Mental Health services	**Children and Young People's services**
· Self-harm · Psychological safety · Whether a victim of violence or aggression · Omissions of medication · Restraint (inpatients only)	· Deterioration · Extravasation · Pain · Skin Integrity

Data for all five NHS Safety Thermometer indicators are collected on a single day each month. They enable wards, teams and organisations to understand the burden of harm to patients (the elderly, women and babies, the mentally ill, children and young people) and of medication error. Data can be used as a baseline to direct improvement efforts and then to measure improvement over time.

5.3 Results of the review

NHS Outcomes Framework Indicators

Tables 6 to 15 summarise the results of the reviewing process. For each indicator-criterion combination, we report the total number of reviewers who answered that the particular indicator satisfies a criterion over the total number of reviewers who have provided an answer. For example,

---

[10] The Maternity Safety Thermometer was first piloted between June 2013 and October 2014 and is now fully released.

[11] The Medication Safety Thermometer was designed to identify harm from medication error in line with the Domain 5 of the NHS Outcomes Framework.

[12] The Mental Health Safety Thermometer was tested in a pilot phase from Nov 2012 until May 2013, and a number of changes were made to the collection over that period of time. A second period of pilot testing and further development ran from the end of April 2014 to October 2014. The official launch of the Mental Health Safety Thermometer took place on the 23rd of October 2014.

in Table 6 only 2 out of 5 reviewers thought that the NHS OF indicator "Life expectancy at 75 – Males" satisfied the "Added value" criterion.

To facilitate the visual identification of indicators that satisfy a certain criterion, indicator-criterion combinations have been colour-coded, with darker blue shades indicating a higher agreement amongst reviewers on whether or not an indicator satisfies a criterion.

NHS OF indicators pertaining to Domain 1 (Table 6) refer to the prevention of premature (avoidable) deaths. Reviewers thought that these did not usually satisfy the "Added value" criterion and that there were problems with "Attribution" to the health care system. The former was found to be the case for almost all indicators, whilst the latter was true for "Life expectancy" both Males and Females and disease specific under 75 mortality rates and excess under 75 mortality rate.

For NHS OF indicators designed to capture the quality of life of individuals with long-term conditions (Domain 2) (Table 7), reviewers agreed that most indicators satisfied all or almost all seven criteria. For five indicator-criterion combinations, a number of exceptions were found as follows: the indicator "Estimated diagnosis rate for people with dementia" was found not to meet the criterion "Benefit/Measure of value"; the indicators "Employment of people with long-term conditions" and "Employment of people with mental illness" did not satisfy the criterion "Attribution"; the indicator "Health-related quality of life for people with long-term conditions" did not satisfy the criteria "Consistency over time" and "Measurability/timeliness"; the indicator "Proportion of people feeling supported to manage their condition" did not meet the criteria "Consistency over time" and "Measurability/timeliness" , and finally the indicator "Health-related quality of life for people with three or more long-term conditions" did not satisfy the criterion "Consistency over time".

For Domain 3 NHS OF indicators (Table 8), most reviewers agreed that almost all indicators met all seven criteria. However, for four indicators for which issues were raised on either their ability to add value, their Consistency over time, or the possibility to attribute the performance measured to the workings of the healthcare system. These were "Emergency readmissions within 30 days of discharge", "Proportion of older people (65 and over) who were still at home 91 days after discharge from hospital into reablement/rehabilitation services" and "Tooth extractions due to decay for children admitted as inpatients to hospital, aged 10 years and under".

NHS OF indicators designed to assess whether patients, service users and carers have a positive experience of care (Domain 4, Table 9) were also found to be either failing to add value or lacking consistency over time. Note that not all reviewers provided an answer as to whether or not the indicators satisfied the "Added value" criterion and, of those who did, about 50% thought that it did so.

Scores for the NHS OF indicators relating to the safe treatment and care of patients (Domain 5) are shown in Table 10. Reviewers felt that these indicators met most criteria, though some questioned whether they "Added value".

Overall, reviewers found for the NHS OF indicators that if these did not meet a criterion, this was most likely due to a failure to satisfy either the "Added value" or the "Consistency over time" criterion. Reviewers found the criteria most likely to be satisfied were "Clarity", "Measurability/timeliness" and "Validity".

Table 6: NHS OF Domain 1: Review of selected criteria set

Domain 1 Preventing people from dying prematurely	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Potential Years of Life Lost (PYLL) from causes considered amenable to healthcare – Adults	6/6	2/4	6/6	4/6	6/6	6/6	6/6
Potential Years of Life Lost (PYLL) from causes considered amenable to healthcare - Children and Young People	6/6	2/4	6/6	4/6	6/6	6/6	6/6
Life expectancy at 75 – Males	6/6	2/5	5/6	1/6	6/6	6/6	4/5
Life expectancy at 75 – Females	6/6	2/5	5/6	1/6	6/6	6/6	4/5
Neonatal mortality and stillbirths	6/6	1/4	6/6	5/6	6/6	6/6	6/6
Under 75 mortality rate from cardiovascular disease	6/6	2/5	6/6	2/5	5/6	6/6	6/6
Under 75 mortality rate from respiratory disease	6/6	2/5	6/6	3/5	5/6	6/6	6/6
Under 75 mortality rate from liver disease	6/6	2/5	6/6	2/5	5/6	6/6	6/6
Under 75 mortality rate from cancer	6/6	1/4	6/6	2/5	5/6	6/6	6/6
One-year survival from all cancers	6/6	0/5	6/6	6/6	5/6	6/6	6/6
Five-year survival from all cancers	6/6	0/4	6/6	6/6	4/5	6/6	6/6
One-year survival from breast, lung and colorectal cancer	6/6	1/5	6/6	6/6	5/6	6/6	6/6
Five-year survival from breast, lung and colorectal cancer	6/6	1/5	6/6	6/6	4/5	6/6	6/6
Excess under 75 mortality rate in adults with serious mental illness	6/6	1/4	6/6	2/5	5/6	5/6	5/6
Infant mortality	6/6	2/5	6/6	4/5	6/6	6/6	6/6
Five-year survival from all cancers in children	6/6	1/4	6/6	5/6	4/5	6/6	6/6

Table 7: NHS OF Domain 2: Review of selected criteria set

Domain 2 Enhancing quality of life for people with long-term conditions	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Employment of people with long-term conditions	6/6	4/5	6/6	0/5	5/6	5/6	5/5
Health-related quality of life for people with long-term conditions	5/6	5/5	6/6	4/5	2/6	2/6	4/6
Unplanned hospitalisation for asthma, diabetes and epilepsy in under 19s	5/6	4/5	4/6	6/6	5/6	6/6	6/6
Proportion of people feeling supported to manage their condition	4/6	3/5	4/6	3/6	2/6	2/6	4/5
Unplanned hospitalisation for chronic ambulatory care sensitive conditions (all ages)	4/6	4/5	4/6	5/6	6/6	6/6	6/6
Employment of people with mental illness	4/6	4/5	5/6	0/5	5/6	5/6	6/6
Estimated diagnosis rate for people with dementia	4/6	3/5	1/6	6/6	4/6	5/6	4/6
Health-related quality of life for people with three or more long-term conditions	4/6	3/5	6/6	4/5	2/6	5/6	4/6
Health-related quality of life for carers	3/6	3/5	5/6	4/6	4/6	4/6	5/6

Table 8: NHS OF Domain 3: Review of selected criteria set

Domain 3

Helping people to recover from episodes of ill health or following injury

Indicator	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Hip fracture: Proportion of patients recovering to their previous levels of mobility at 30 days	6/6	3/4	6/6	6/6	3/6	4/6	6/6
Hip fracture: Proportion of patients recovering to their previous levels of mobility at 120 days	6/6	4/4	6/6	4/6	3/6	5/6	6/6
Emergency readmissions within 30 days of discharge from hospital	5/6	5/5	6/6	6/6	6/6	6/6	6/6
Emergency admissions for children with lower respiratory tract infections	5/6	2/4	4/6	4/6	6/6	6/6	6/6
Proportion of older people (65 and over) who were still at home 91 days after discharge from hospital into reablement/rehabilitation services	5/6	3/4	6/6	3/6	3/6	5/6	4/6
Proportion offered rehabilitation following discharge from acute or community hospital	5/6	2/4	5/6	4/6	2/6	5/6	5/6
Tooth extractions due to decay for children admitted as inpatients to hospital, aged 10 years and under	5/6	3/4	4/5	2/5	5/6	6/6	5/6
Emergency admissions for acute conditions that should not usually require hospital admission	4/6	3/4	5/6	5/6	6/6	6/6	5/6

Table 9: NHS OF Domain 4: Review of selected criteria set

Domain 4

Ensuring that people have a positive experience of care

Indicator	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
GP out-of-hours services	6/6	3/4	6/6	6/6	2/6	5/6	5/6
Patient experience of hospital care	6/6	2/3	5/6	5/6	5/6	5/6	5/6
Patient experience of outpatient services	6/6	2/4	6/6	6/6	1/6	5/6	5/6
Patient experience of A&E services	6/6	2/4	5/6	6/6	0/6	5/6	6/6
Patient experience of primary care - GP services	5/6	2/4	6/6	6/6	3/6	5/6	5/6
Responsiveness to inpatients' personal needs	5/6	1/3	5/6	3/6	3/4	3/6	2/3
Access to GP services	5/6	2/4	6/6	6/6	2/6	5/6	6/6
Access to NHS dental services	5/6	2/4	6/6	5/6	2/6	5/6	6/6
Women's experience of maternity services	5/6	2/4	6/6	5/6	0/6	5/6	5/6
Bereaved carers' views on the quality of care in the last 3 months of life	5/6	2/4	6/6	3/6	0/6	2/5	2/6
Patient experience of community mental health services	5/6	2/4	5/6	5/6	1/4	5/6	4/5
NHS dental services	3/6	2/4	5/6	5/6	2/4	3/4	3/4

Table 10: NHS OF Domain 5: Review of selected criteria set

Domain 5

Treating and caring for people in a safe environment and protecting them from avoidable harm

Indicator	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Deaths from VTE related events within 90 days post discharge from hospital	6/6	4/4	6/6	3/5	6/6	6/6	6/6
Incidence of healthcare-associated infection - MRSA bacteraemia	6/6	3/4	6/6	6/6	6/6	6/6	6/6
Incidence of healthcare-associated infection - C.difficile	6/6	3/4	6/6	6/6	6/6	6/6	6/6
Admission of full-term babies to neonatal care	5/6	2/4	5/6	5/6	4/6	6/6	4/6
Patient safety incidents reported	4/6	2/4	5/6	5/6	1/6	3/6	2/6

NHS Thermometer Indicators

The first set of NHS Safety Thermometer indicators in Table 11 relate to the 'Classic' thermometer. The majority of reviewers found that this set of indicators did not satisfy two criteria: "Added value" and "Consistency over time", whilst agreeing, in some cases unanimously that they satisfied the remaining criteria.

Regarding the indicators measuring harm caused by medication errors (Table 12), all six reviewers answered that the indicators satisfied the "Clarity", "Benefit/Measure of value" and the "Attribution" criteria. Only one reviewer indicated that these indicators satisfied the "Consistency over time" criterion. Similarly, for the "Added value" criterion, reviewers were generally satisfied that the Medication Safety indicators satisfied this particular criterion. Regarding both the "Measurability/timeliness" and "Validity" criteria, the majority of respondents agreed that the criteria were met by the indicators.

Reviewers in general did not find that the NHS Mental Health Safety Thermometer indicators (Table 13) satisfied the "Added value", "Attribution" (except for the indicators 'Proportion of patients that have had an omission of medication in the last 24 hours' and 'Proportion of patients that have been restrained in the last 72 hours" and the "Consistency over time" criteria. For these criteria, the reviewers either did not provide an answer at all or did not find that a particular indicator met the criterion. For the remaining criteria, the majority of reviewers agreed that they were met by the indicators.

NHS Safety Thermometer indicators assessing the safety of healthcare services delivered to maternity care patients (mothers and babies) were found to satisfy in general all criteria (some with unanimous agreement amongst reviewers), except for the criterion "Consistency over time" for which only one reviewer found them to satisfy it.

A similar outcome to the one that emerged from the reviewing process for the Maternity services appeared for the last set of NHS Safety Thermometer indicators, namely the Children and Young People indicators.

Overall, we found general agreement among reviewers that NHS Thermometer indicators were most likely to be excluded from further consideration because reviewers felt that the indicator either failed to add value or lacked consistency over time. Reviewers found that most indicators met the remaining criteria.

Table 11: NHS Safety Thermometer - Classic: Review of selected criteria set

Classic							
	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Pressure Ulcers	4/6	1/3	6/6	6/6	1/5	5/6	6/6
Falls	4/6	1/3	6/6	5/6	1/5	5/6	5/6
VTE	5/6	0/3	6/6	5/6	1/5	5/6	6/6
Catheters	5/6	1/3	6/6	6/6	1/5	5/6	6/6

Table 12: NHS Safety Thermometer - Medication: Review of selected criteria set

Medication							
	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Proportion of patients with reconciliation started within 24 hours of admission	6/6	2/4	6/6	6/6	1/4	5/6	5/5
Proportion of patients who have had an omitted dose in the last 24 hours	6/6	2/5	6/6	6/6	1/5	5/6	5/5
Proportion of patients with medicine allergy status documented	6/6	3/5	6/6	6/6	1/4	5/6	5/5
Proportion of patients with an omission of a critical medicine	6/6	3/5	6/6	6/6	1/5	5/6	5/5
Proportion of patients receiving a high risk medication in the last 24 hours	6/6	3/5	6/6	6/6	1/4	5/6	5/6
Proportion of patients on a high risk medicine that trigger an MDT referral	6/6	3/5	6/6	6/6	1/3	6/6	4/4

Table 13: NHS Safety Thermometer – Mental Health: Review of selected criteria set

Mental health	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Proportion of patients that have self harmed in the last 72 hours	5/6	1/4	6/6	0/4	1/5	5/6	4/5
Proportion of patients that feel safe at the point of survey	6/6	1/4	5/6	0/4	1/5	5/6	4/5
Proportion of patients that have been the victim of violence/aggression (last 72 hours)	5/6	1/4	5/6	0/4	1/5	5/6	4/5
Proportion of patients that have had an omission of medication in the last 24 hours	6/6	2/4	6/6	4/5	1/5	5/6	5/6
Proportion of patients that have been restrained in the last 72 hours	4/6	1/4	5/6	3/5	1/4	6/6	3/4

Table 14: NHS Safety Thermometer - Maternity: Review of selected criteria set

Maternity	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Proportion of women that had a maternal infection	6/6	2/5	6/6	5/6	1/5	5/6	5/6
Proportion of women that had a 3rd/4th degree perineal trauma	6/6	4/5	6/6	6/6	1/5	5/6	6/6
Proportion of women that had a PPH of more than 1000mls	6/6	4/5	6/6	6/6	1/5	5/6	6/6
Proportion of Term Babies with an Apgar less than 7 at 5 Minutes	6/6	4/5	6/6	4/5	1/5	5/6	6/6
Proportion of women who were left alone at a time that worried them	6/6	3/5	6/6	6/6	1/5	5/6	6/6
Proportion of women with concerns about safety during labour and birth not taken seriously	5/6	3/5	6/6	6/6	1/5	5/6	6/6

Table 15: NHS Safety Thermometer – Children and Young People: Review of selected criteria set

Children and Young People	Clarity	Added value	Benefit/Measure of value	Attribution/ granularity	Consistency over time	Measurability/ timeliness	Validity
Indicator							
Proportion of patients with an EWS not completed for each set of observations in the last 12 hours	5/6	2/5	5/6	5/6	1/5	5/6	5/6
Proportion of patients with an EWS completed, triggered and not escalated	5/6	2/5	5/6	5/6	1/5	5/6	5/6
Proportion of patients with extravasation in the last 24 hours	6/6	2/5	6/6	6/6	1/5	5/6	6/6
Proportion of patients in pain at the point of survey	6/6	2/5	6/6	6/6	1/5	5/6	6/6
Proportion of patients with a pressure ulcer (new or old)	6/6	3/5	6/6	6/6	1/5	5/6	6/6
Proportion of patients with a moisture lesion (new or old)	6/6	3/5	6/6	6/6	1/5	5/6	6/6

5.4 Which indicators met most criteria?

We draw the preceding material to identify those indicators that satisfy the greatest number of criteria. In performing this assessment, the seven criteria are afforded equal weight. This simplifies the task but also reflects the views expressed by UK experts at the workshop.

Given that reviewers expressed different views on whether indicators satisfied each criterion, we set thresholds about how much agreement there had to be among reviewers. The strictest threshold required that, for an indicator to be considered, 80% of the reviewers had to have said that all criteria were satisfied by the indicator in question. We also report how many indicators would be selected if the threshold was progressively relaxed to 70%, 60% and 50% consensus levels.

Only one NHS Outcomes Framework indicator met the strictest cut-off rule (80% of consensus), three indicators met the 70% cut-off rule, 8 indicators met the 60% cut-off rule and finally 17 indicators met the 50% cut-off rule. A summary of these indicators can be found in Appendix A.

Of these 17, two NHS OF indicators pertain to Domain 1, three indicators pertain to Domain 2, five indicators pertain to Domain 3, three indicators pertain to Domain 4 and six indicators pertain to Domain 5. We show the selected indicators in the Venn-Diagram in Figure 2.

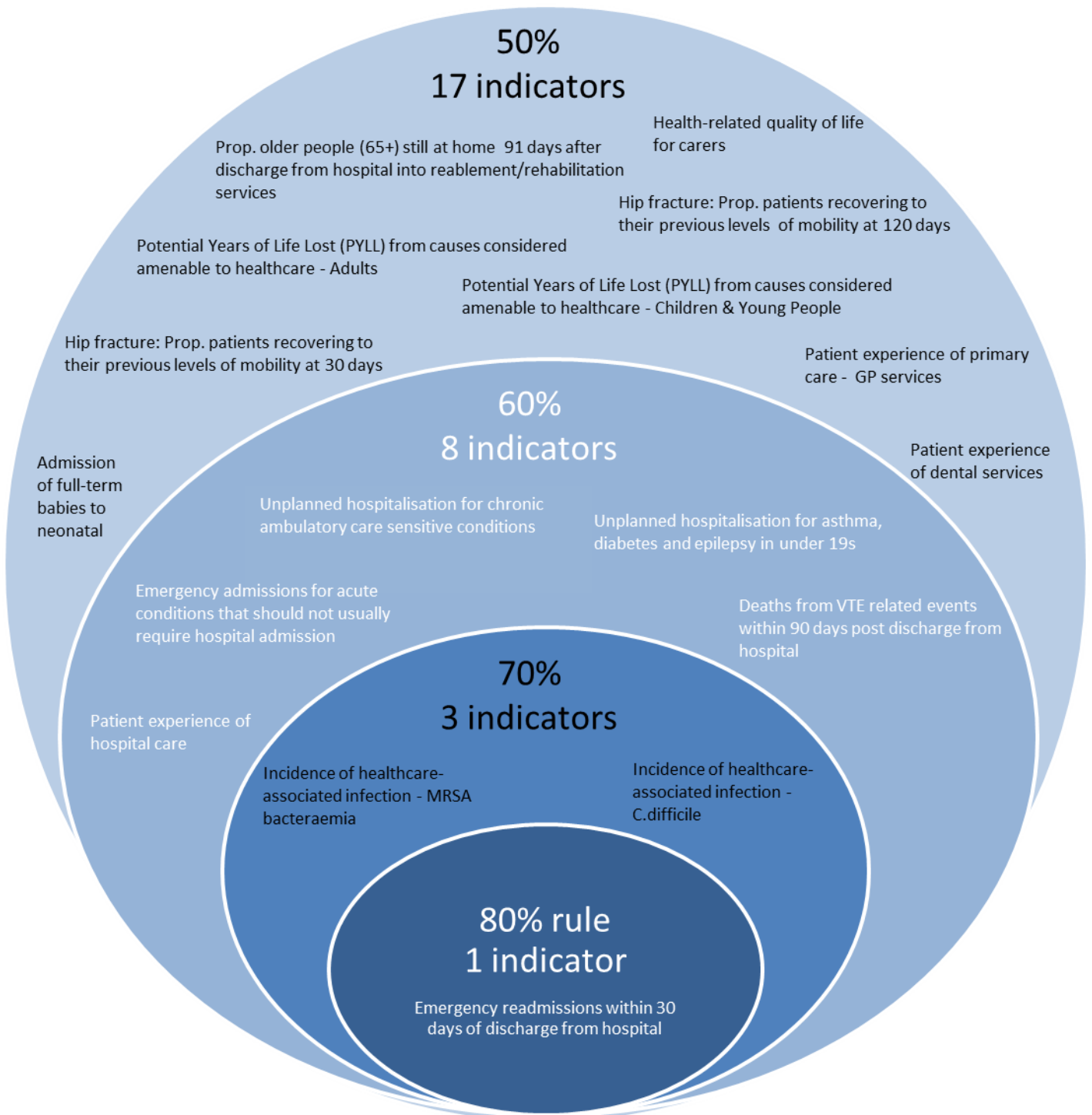No NHS Thermometer indicators could be selected based on any of the above cut-off rules.

Figure 2: Venn Diagram of selected NHS Outcomes Framework indicators based on different cut-off rules

6. Conclusions

A measure of the output of the health system should capture the value of the output produced. This is challenging because people usually do not demand healthcare for its own sake, but because of its contribution towards improving their health. Capturing the value of healthcare therefore requires some means of measuring both the outcome and process of healthcare delivery.

Current practice in accounting for the quality of healthcare services makes use of routinely available information in order to capture the QALYs associated with treating patients, by combining information on survival rates, life expectancy and a measure of change in health status before and after treatment. The process of care delivery is captured by measures of treatment waiting times.

This approach may overlook other important characteristics of the quality of healthcare. This review was designed to assess whether other indicators of quality could be incorporated into the NHS output measure. To this end, the review aimed first to identify which criteria an indicator ought to satisfy in order to be considered for inclusion in an output measure.

To this end we evaluated three published sets of criteria developed by national bodies responsible for assessing healthcare system performance. We sought the opinions of UK experts on quality expressed at the workshop. From this process seven criteria were established, these being:

- Clarity
- Added value
- Benefit / measure of value
- Attribution / granularity
- Consistency over time
- Measurability / timeliness
- Validity

We next reviewed two sources of quality indicators currently collected and reported for the English NHS: the NHS Outcomes Framework indicators and NHS Thermometer indicators. A schema, including indicator name and source, data source, time period covered, definitions and purpose, was developed for each of the indicators. Each indicator was assessed by the research team, and one expert from the Department of Health and one from the Office for National Statistics in order to establish whether they met each of the identified criteria. Depending on the level of consensus among reviewers, a maximum of 17 indicators were short-listed for potential use as quality adjustors for NHS output, all of which are NHS Outcomes Framework indicators. These potential indicators, ordered according to those best judged to have met the criteria, were:

Threshold	Indicator
80%	Emergency readmissions within 30 days of discharge from hospital
70%	Incidence of healthcare associated infection – MRSA bacteraemia; Incidence of healthcare associated infection – C.difficile
60%	Unplanned hospitalisation for chronic ambulatory care sensitive conditions; Unplanned hospitalisation for asthma, diabetes and epilepsy in under 19s; Emergency admissions for acute conditions that should not usually require hospital admission; Patient experience of hospital care; Deaths from venous thromboembolism (VTE) related events within 90 days post discharge from hospital
50%	Potential Years of Life Lost (PYLL) from causes considered amenable to healthcare – Adults; Potential Years of Life Lost (PYLL) from causes considered amenable to healthcare – Children and Young people; Health-related quality of life for carers; Hip fracture: proportion of patients recovering to their previous levels of mobility / walking ability at 30 days; Hip fracture: proportion of patients recovering to their previous levels of mobility / walking ability at 120 days; Proportion of older people (65 and over) who were still at home 91 days after discharge from hospital into reablement/rehabilitation services; Patient experience of primary care – GP services; Patient experience of dental services; Admission of full-term babies to neonatal care

Further research is needed to establish how best to incorporate the set of indicators in the NHS output measure. The main challenge will be to establish what weight (value) should be attached to each quality dimension so that it can be incorporated in the overall output measure.

Overall, we recommend that sensitivity analyses are performed for the selected indicators in order to establish whether their effects on the NHS output growth measure are in line with prior expectations as well as to ensure consistency of the NHS output measure over time.

References

1.   Castelli, A., et al., The challenges of measuring government output in the health care sector. *Social Policy Review* 2005;**17**:p.183-202.
2.   Bojke, C., et al., Productivity growth in the English National Health Service from 1998/1999 to 2013/2014. *Health Economics* DOI:10.1002/hec.3338, 2016.
3.   Dawson, D., et al., Developing new approaches to measuring NHS outputs and productivity, Final Report. 2005, York: *Centre for Health Economic Research Paper 6*.
4.   Black, N., Declining health-care productivity in England: the making of a myth. *The Lancet;* **379**(9821):p. 1167-1169.
5.   Atkinson, T., *Measuring health output, productivity and equity.* 2010, London: Office of Health Economics.
6.   National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal 2013. 2013 18 December 2015]; Available from: https://www.nice.org.uk/article/pmg9/chapter/foreword.
7.   Smith, P.C. and A. Street, Using patient reported outcome measures. *Health Economics* 2013;**22**(2): p. 119-131.
8.   Castelli, A., et al., A new approach to measuring health system output and productivity. *National Institute Economic Review* 2007;**200**:p. 105-117.
9.   Department of Health, Guidance on the routine collection of patient reported outcome measures (PROMs). 2007, London: Department of Health.
10.   Office for National Statistics. National life tables, UK Statistical bulletins. 2017 [cited 2017 10 July 2017]; Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/previousReleases.
11.   Derbyshire, K., et al., Further developments in measuring quality adjusted healthcare output. 2007, Leeds: Department of Health.
12.   Bojke, C., et al., Productivity Of The English NHS: 2013/14 update, in CHE Research Paper 126. 2016, Centre for Health Economics, University of York: York. p. 69.
13.   Dawson, D., et al., Developing New Approaches to Measuring NHS Outputs and Activity, in CHE Research Paper 6. 2005, Centre for Health Economics, University of York: York.
14.   Claxton, K., et al., Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technol Assessment* 2015;**19**(14).
15.   Currie, C.J., et al., The Routine Collation of Health Outcomes Data from Hospital Treated Subjects in the Health Outcomes Data Repository (HODaR): Descriptive Analysis from the First 20,000 Subjects. *Value in Health* 2005;**8**(5): p. 581-590.
16.   Agency for Healthcare Research and Quality and M. Farquhar, AHRQ Quality Indicators, in Quality Indicators Patient Safety and Quality: An Evidence-Based Handbook for Nurses, R.G. Hughes, Editor. 2008, Agency for Healthcare Research and Quality: Rockville (MD).
17.   Agency for Healthcare Research and Quality, Refinement of the HCUP quality indicators. Technical Review No. 4 (AHRQ publication No. 01-0035). 2001, Rockville (MD): Agency for Healthcare Research and Quality.
18.   National Centre for Health Outcomes Development, Annex 12: Evaluating the Quality of Clinical and Health Indicators, in Compendium of Clinical and Health Indicators User Guide. 2005.
19.   HSCIC (Indicator Assurance Service), Criteria and considerations used to determine a quality indicator. Vol. Version 3. 2014, Leeds: Health and Social Care Information Centre.

Appendix A: Summary of selected NHS Outcomes Framework indicators

Domain 1 Overreaching Indicators	
Indicator:	Potential Years of Life Lost (PYLL) from causes considered amenable to healthcare - Adults
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data source:	ONS avoidable mortality / ONS populations / ONS period and cohort life expectancy
Years covered:	2009-2014
Recent changes:	N/A

Definition

The number of years of life lost by every 100,000 adults aged 20 and over dying from a condition which is usually treatable, measured in a way which allows for comparisons between populations with different age profiles and over time.

Purpose

To ensure that the NHS is held to account for doing all that it can to prevent amenable deaths. Deaths from causes considered 'amenable' to healthcare are premature deaths that should not occur in the presence of timely and effective healthcare.
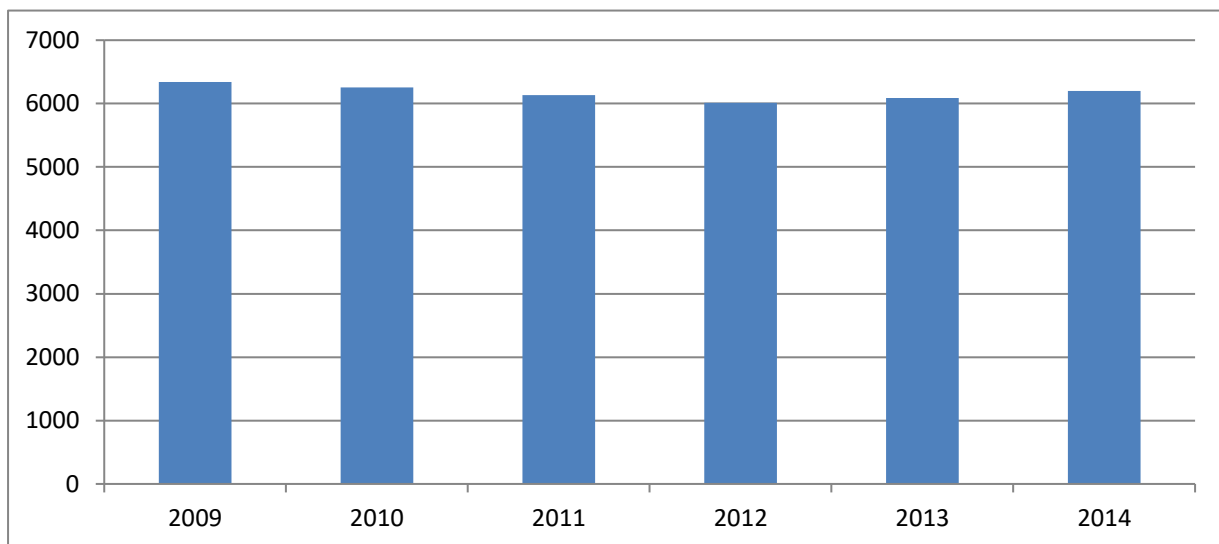

Figure A1: Directly age and sex standardised potential years of life lost (PYLL) per 100,000 registered patients, 95% confidence intervals (CI)

Comments:

Although the values change over time, the absolute difference is very small and very unlikely to make any difference.

Domain 1	
Overreaching Indicators	
Indicator:	Potential years of life lost (PYLL) from causes considered amenable to healthcare - children and young people
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data source:	ONS avoidable mortality / ONS populations / ONS period and cohort life expectancy
Years covered:	2003-2014
Recent changes:	Retired in the 2015/16 NHS Outcomes Framework

Definition

The number of years of life lost by every 100,000 persons aged 0 to 19 dying from a condition which is usually treatable, measured in a way which allows for comparisons between populations with different age profiles and over time.

Purpose

To ensure that the NHS is held to account for doing all that it can to prevent amenable deaths. Deaths from causes considered 'amenable' to healthcare are premature deaths that should not occur in the presence of timely and effective healthcare.
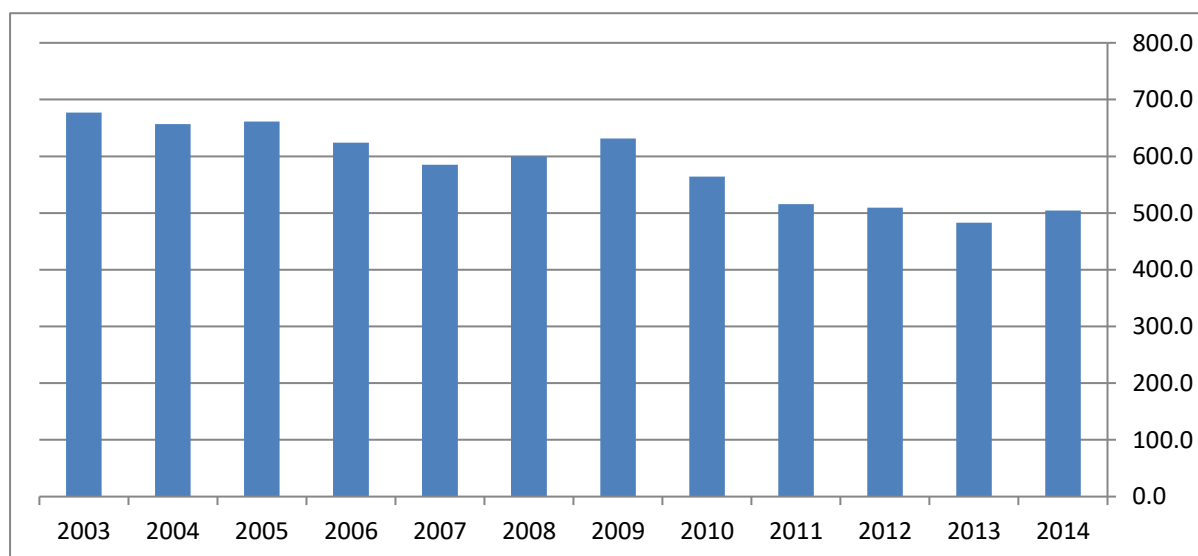
Figure A2: Potential years of life lost (PYLL) from causes considered amenable to healthcare - children and young people

Comments:

Bigger variation than for adults, but in absolute number the difference will not make an impact on our index.

Domain 2
Improvement areas: Reducing time spent in hospital by people with long-term conditions

Indicator:	Unplanned hospitalisation for chronic ambulatory care sensitive conditions
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data source:	HES / ONS populations
Years covered:	2003/04-2015/16
Recent changes:	N/A

Definition

This indicator measures how many people with specific long-term conditions, which should not normally require hospitalisation, are admitted to hospital in an emergency. These conditions include, for example, diabetes, epilepsy and high blood pressure.

Purpose

This outcome is concerned with how successfully the NHS manages to reduce emergency admissions for all long-term conditions where optimum management can be achieved in the community.
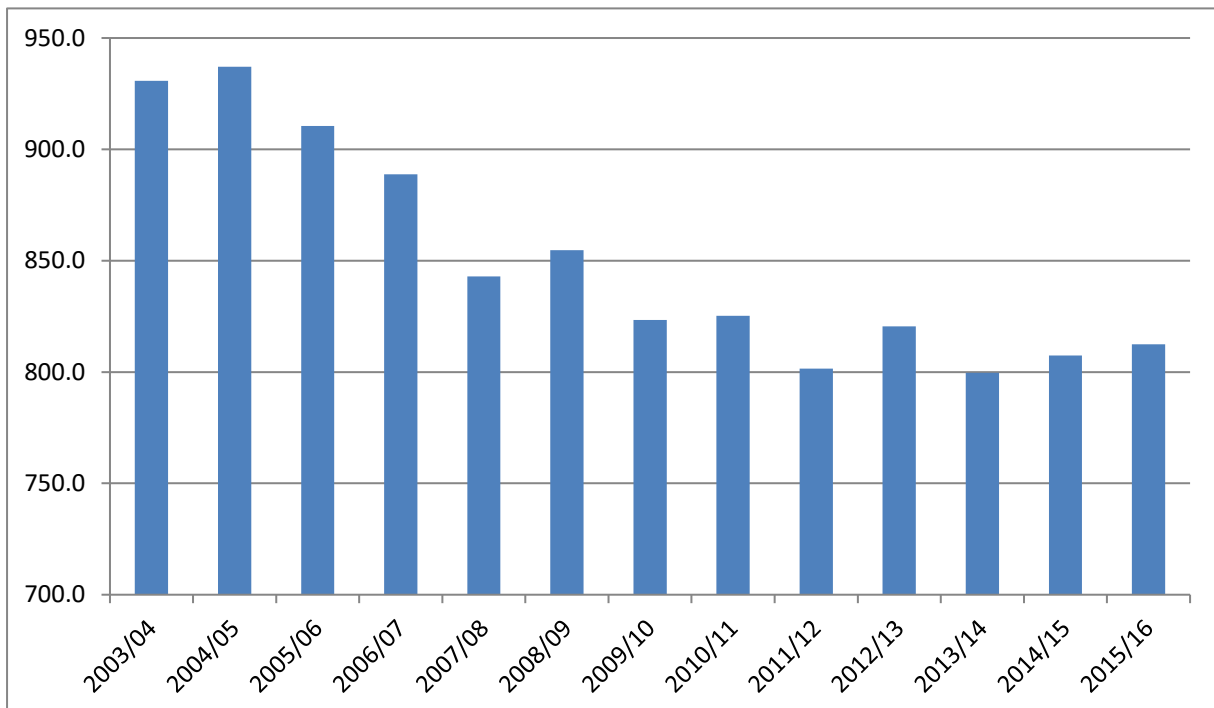
Figure A3: Directly age and sex standardised rate of unplanned hospital admissions

Comments

The indicator has changed over the years it has been measured, but the year-on-year differences, especially in the last financial years, are small.

Domain 2	
Improvement areas: Reducing time spent in hospital by people with long-term conditions	
Indicator:	Unplanned hospitalisation for asthma, diabetes and epilepsy in under 19s
Source:	NHS Outcomes Framework
	(last accessed8 Jun 2017)
Data source:	HES / ONS populations
Years covered:	2003/04-2015/16
Recent changes:	N/A

Definition

This indicator measures how many young people (aged 0-18 inclusive) who have asthma, diabetes or epilepsy are admitted to hospital in an emergency.

Purpose

This outcome is concerned with how successfully the NHS manages to reduce avoidable emergency admissions for children with asthma, diabetes or epilepsy. These three conditions were chosen as they account for around 94 per cent of emergency admissions for children (under 19s) with long-term conditions.
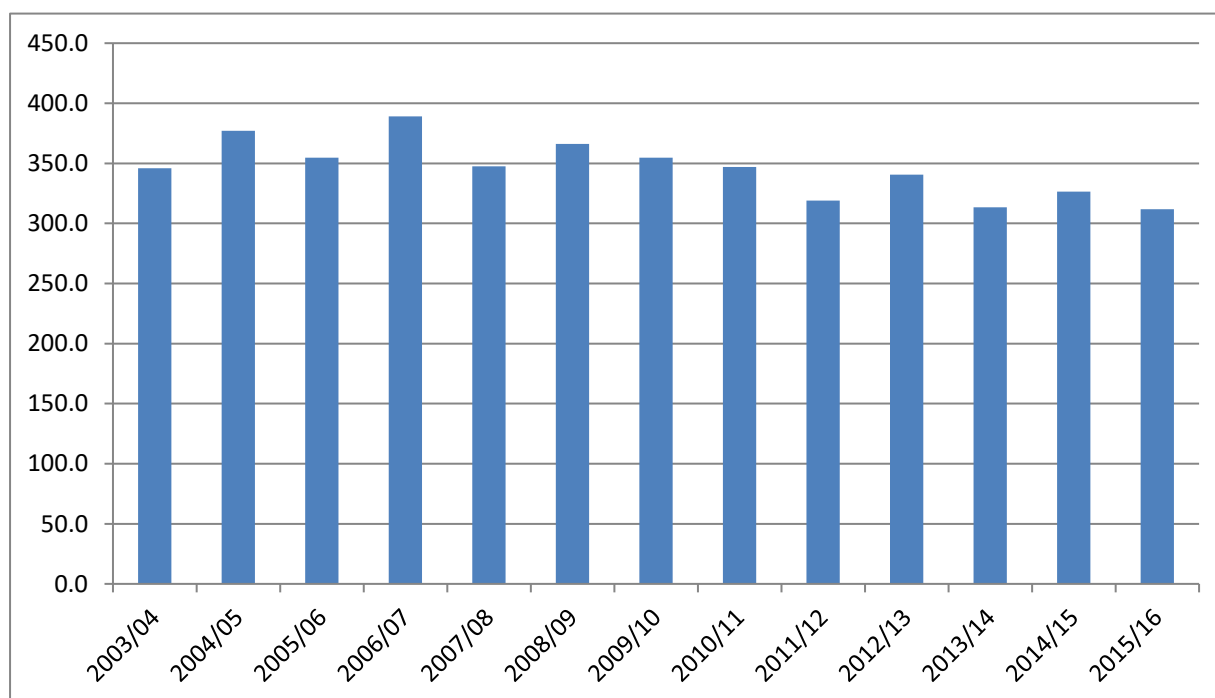
Figure A4: Directly age and sex standardised rate of unplanned hospital admissions for asthma, diabetes and epilepsy in under 19s

Comments

The indicator has changed over the years it has been measured, but the year-on-year differences, especially in the last financial years, are small.

Domain 2	
Improvement areas: Enhancing quality of life for carers	
Indicator:	Health-related quality of life for carers
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data source:	GPPS
Years covered:	2011/12-2015/16 (July to March)
Recent changes:	N/A

Definition

This indicator measures health-related quality of life for people who identify themselves as helping or supporting family members, friends, neighbours or others with their long-term physical or mental illness/disability or because of problems related to old age.

Purpose

This indicator seeks to capture how successfully the NHS is supporting carers to live as normal a life as possible. This indicator helps people understand whether health-related quality of life for carers is improving over time.
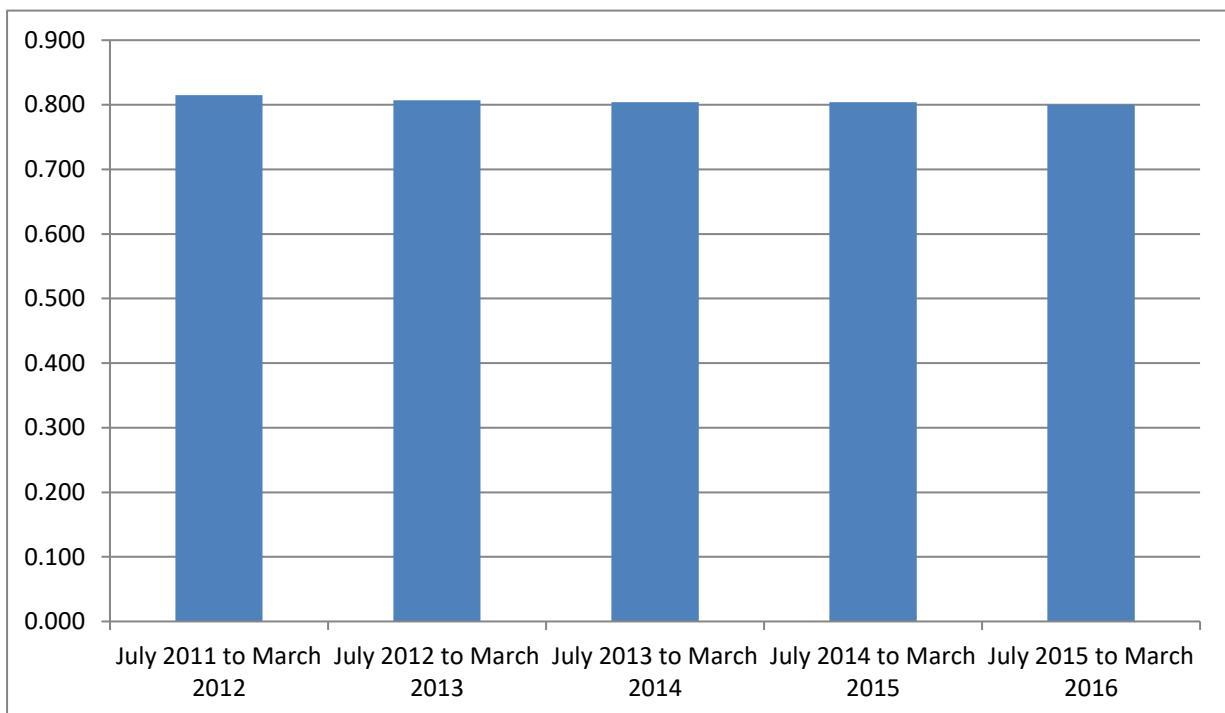
Figure A5: Average adjusted health status (EQ-5D™) score for individuals reporting that they are carers, aged 18 and above, based on responses to a question from the GP Patient Survey

Comments

No variation over time, the indicator has practically constant values. Informal care might come under social services as opposed to the NHS.

| **Domain 3** |
| **Improvement areas: Improving recovery from fragility fractures** |

Indicator:	Hip fracture: Proportion of patients recovering to their previous levels of mobility / walking ability at 30 days
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	The National Hip Fracture Database (NHFD))
Years covered:	2011 to 2015 (calendar year)
Recent changes:	N/A

Definition

This indicator measures the proportion of patients, expressed as a percentage, with a hip fracture recovering to their previous levels of mobility at 30 days.

Purpose

This indicator aims to measure how effectively people recover their mobility following a hip fracture.
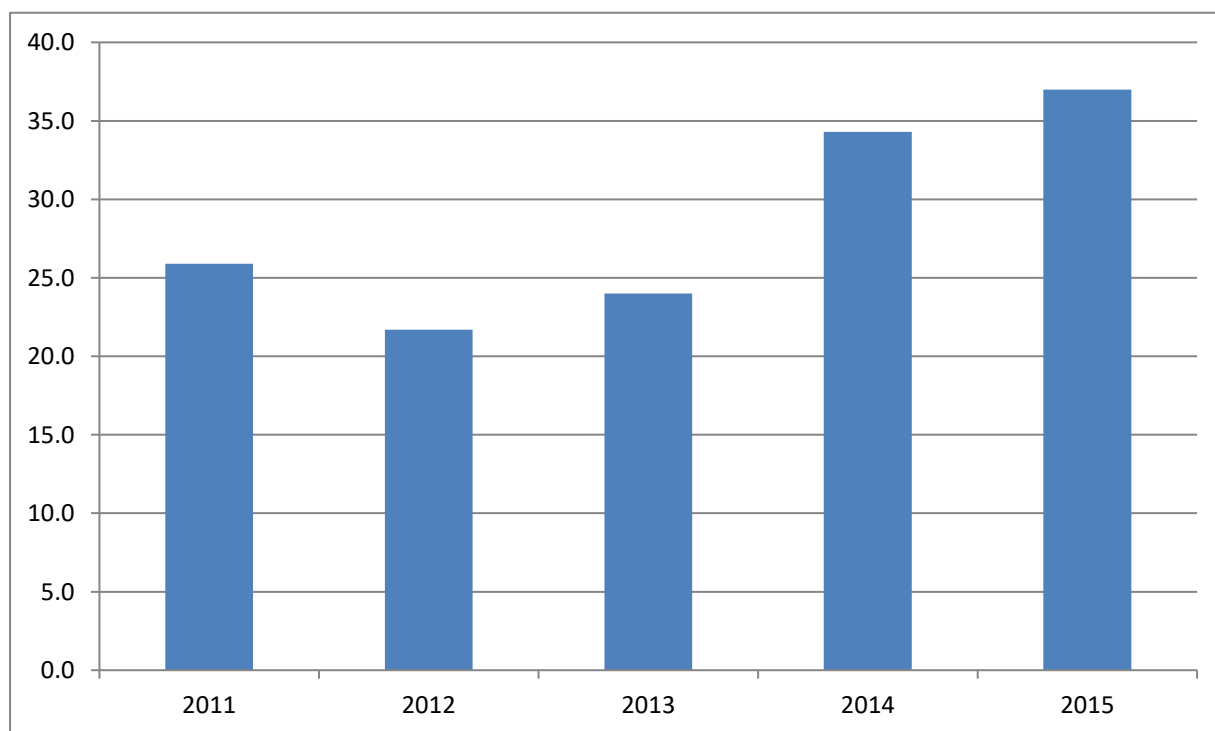
Figure A6: Proportion of patients recovering to their previous levels of mobility/walking ability at 30 days

Comments

Value of indicator has increased significantly between 2013 and 2015. Shortcoming: only four data points are available.

Domain 3
Improvement areas: Improving recovery from fragility fractures

Indicator:	Hip fracture: Proportion of patients recovering to their previous levels of mobility / walking ability at 120 days
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	The National Hip Fracture Database (NHFD))
Years covered:	2011 to 2015 (calendar year)
Recent changes:	N/A

Definition

This indicator measures proportion of patients, expressed as a percentage, with a hip fracture recovering to their previous levels of mobility at 120 days.

Purpose

This indicator aims to measure how effectively people recover their mobility following a hip fracture.
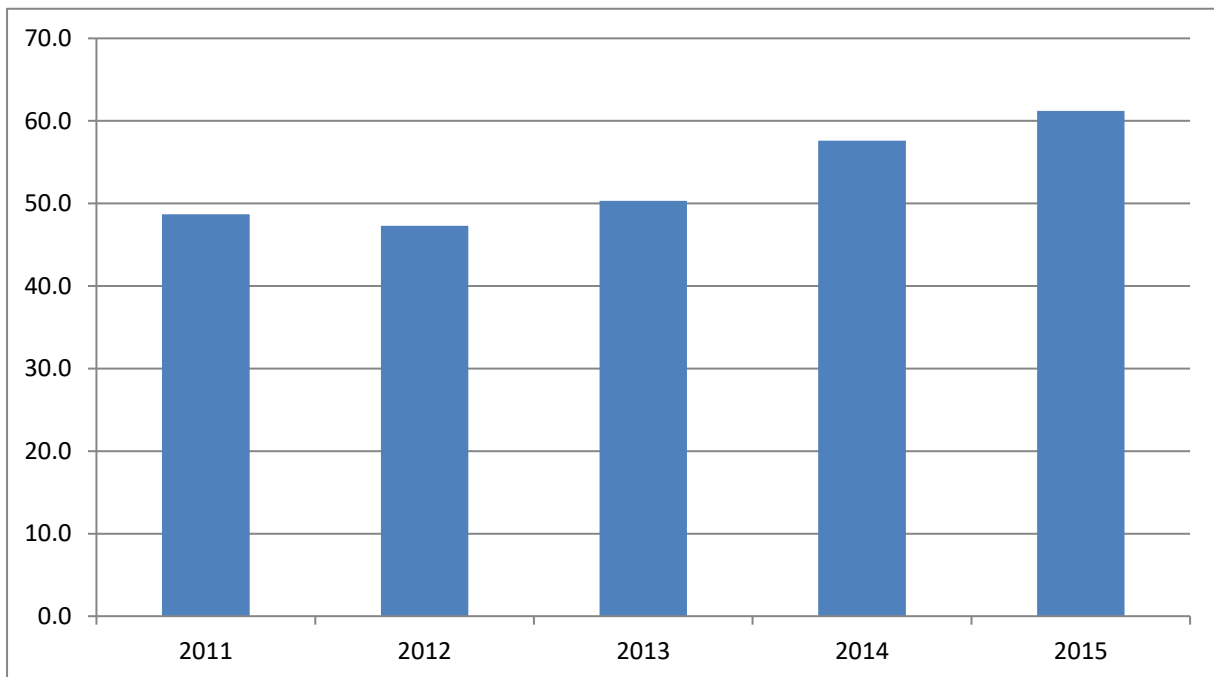
Figure A7: Proportion of patients recovering to their previous levels of mobility/walking ability at 120 days

Comments

Value of indicator has been increased significantly (10.3 percentage points) between 2013 and 2015. Shortcoming: only four data points are available.

Domain 3	
Overreaching Indicators	
Indicator:	Emergency readmissions within 30 days of discharge from hospital
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data Source:	Hospital Episode Statistics (HES), Admitted Patient Care (APC)
Years covered:	2002/03 – 2011/12
Recent changes:	N/A

Definition

This indicator measures the percentage of admissions of people who returned to hospital as an emergency within 30 days of the last time they left hospital after a stay. Admissions for cancer and obstetrics are excluded as they may be part of the patient's care plan.

Purpose

This indicator aims to measure the success of the NHS in helping people to recover effectively from illnesses or injuries. If a person does not recover well, it is more likely that they will require hospital treatment again within the 30 days following their previous admission. Thus, readmissions are widely used as an indicator of the success of healthcare in helping people to recover.
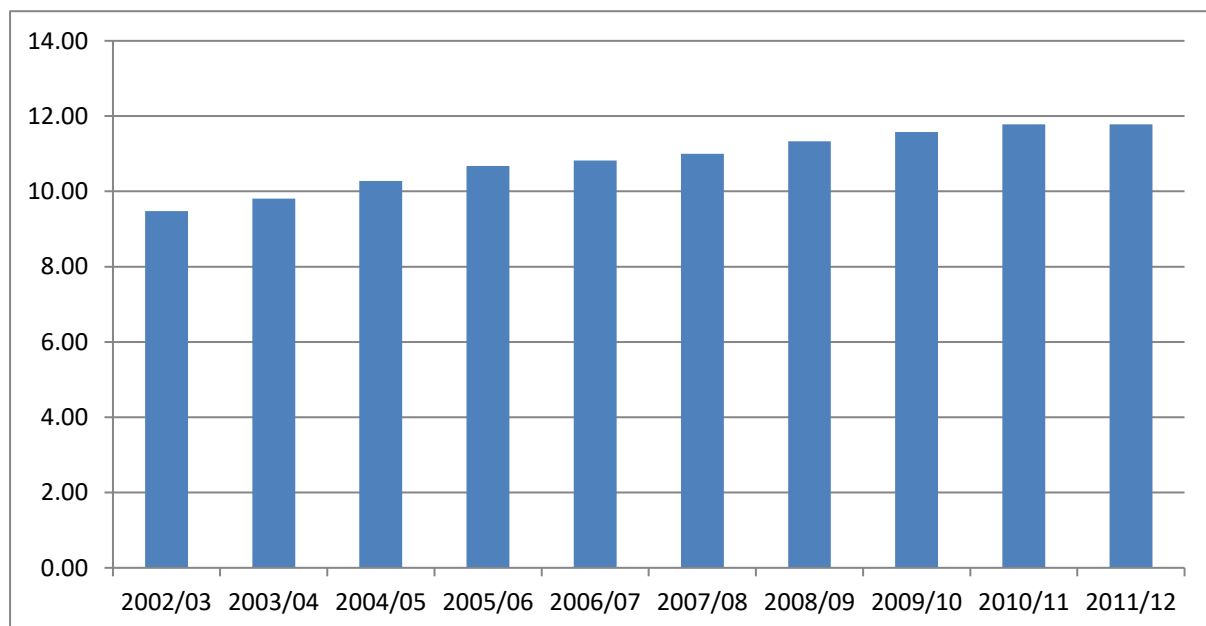

Figure A8: Percentage of Emergency admissions occurring within 30 days of last previous discharge from hospital. Indirectly standardised rate (excl. cancer and obstetrics), 95% confidence intervals (CI)

Comments

Indicator has gradually increased over the time series. The overall increase between 2002/03 and 2011/12 is equal to 24.3 percentage points.

The indicator is currently on hold due to a methodological review.

Domain 3	
Overreaching Indicators	
Indicator:	Emergency admissions for acute conditions that should not usually require hospital admission
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data source:	Hospital Episode Statistics (HES), Admitted Patient Care (APC)
Years covered:	2003/04-2015/16
Recent changes:	N/A

Definition

The indicator measures the number of emergency admissions to hospital in England for acute conditions such as ear/nose/throat infections, kidney/urinary tract infections and heart failure, among others, that could potentially have been avoided if the patient had been better managed in primary care.

Purpose

This indicator aims to measure the reduction in emergency admissions for conditions that should usually be managed outside hospital. Where an individual has been admitted for one of these conditions, it may indicate that they have deteriorated more than should have been allowed by the adequate provision of healthcare in primary care or as a hospital outpatient.
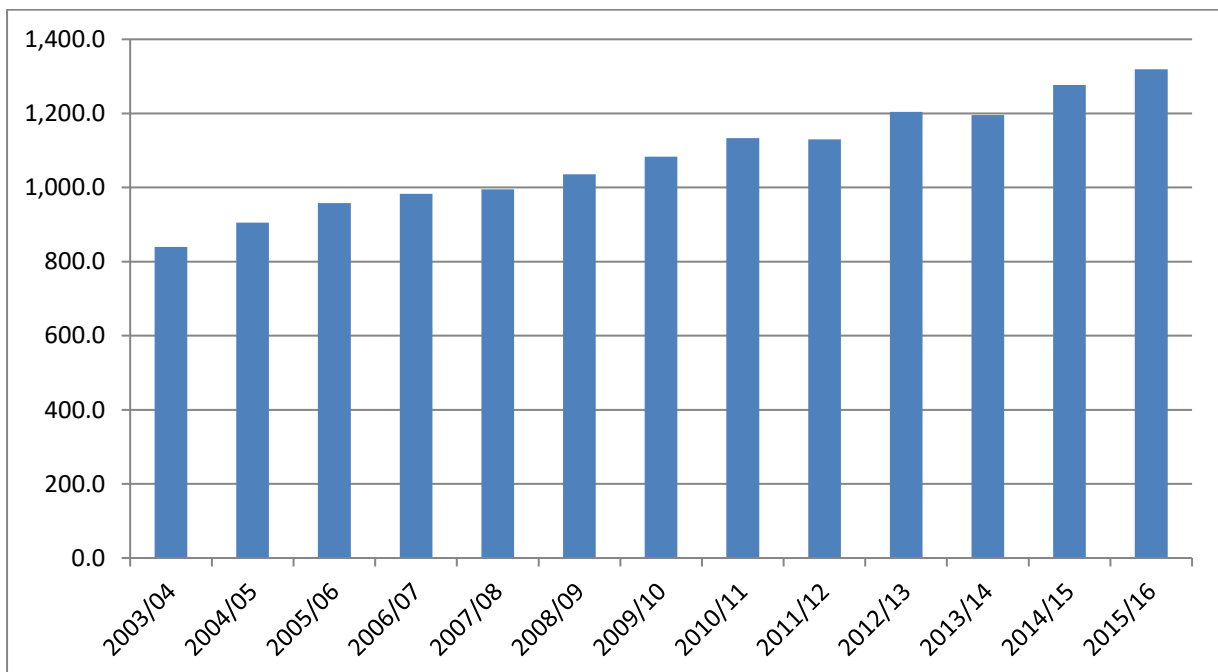

Figure A9: Emergency admissions for acute conditions that should not usually require hospital admission

Comments

Indicator has increased significantly over the time period considered. The overall increase in this indicator of emergency admissions is just under 6 per cent between 2003/04 and 2015/16.

Domain 3	
Improvement areas: Helping older people to recover their independence after illness or injury	
Indicator:	Proportion of older people (65 and over) who were still at home 91 days after discharge from hospital into reablement/rehabilitation services
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	Short and Long-Term Data Return (SALT) and Hospital Episode Statistics (HES)
Years covered:	2011/12-2015/16 (financial years? – not indicated)
Recent changes:	N/A

Definition

This indicator measures the proportion of older people (aged 65 and over), expressed as a percentage, who, after a period of reablement/rehabilitation, maintain their independence by remaining or returning to their home or previous residence 91 days after leaving hospital.

Purpose

This indicator measures the benefit to individuals from reablement, intermediate care and rehabilitation following a hospital episode, by determining whether an individual remains living at home 91 days following discharge. This is seen as a key outcome for many people using reablement services.
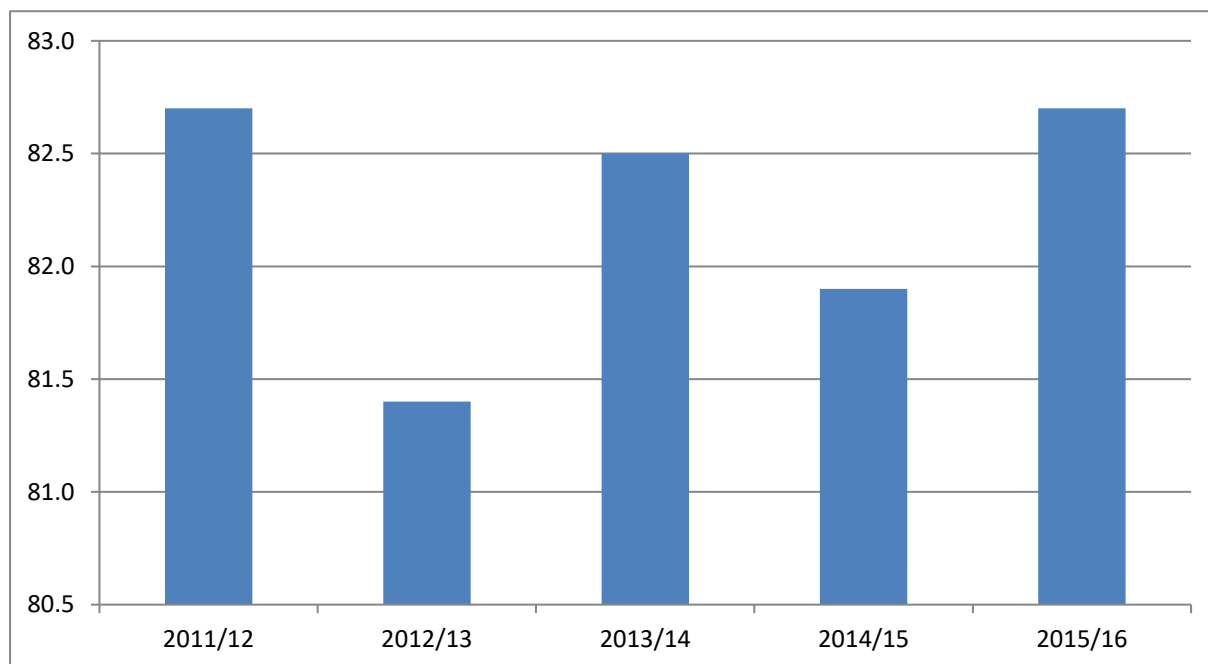

Figure A10: Proportion of older people (65 and over) who were still at home 91 days after discharge from hospital into reablement/rehabilitation services Add figure and text

Comments

Very small change over the four year period. Shortcoming: only five data points are available.

Domain 4	
Overreaching Indicators	
Indicator:	Patient experience of GP services
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	GP Patient Survey (GPPS) from Ipsos MORI (http://www.gp-patient.co.uk) – Official Statistics
Years covered:	2011/12-2015/16 (financial year – period coverage July to September of previous year and January to March of following year for the respective financial year)
Recent changes:	N/A

Definition
This indicator measures the weighted percentage of people who report their overall experience of GP services as 'fairly good' or 'very good'.

Purpose
This indicator aims to capture the experience of patients of their GP. The vast majority of the population visit their GP each year and often it is the experience people have of primary care that determines their overall view of the NHS.
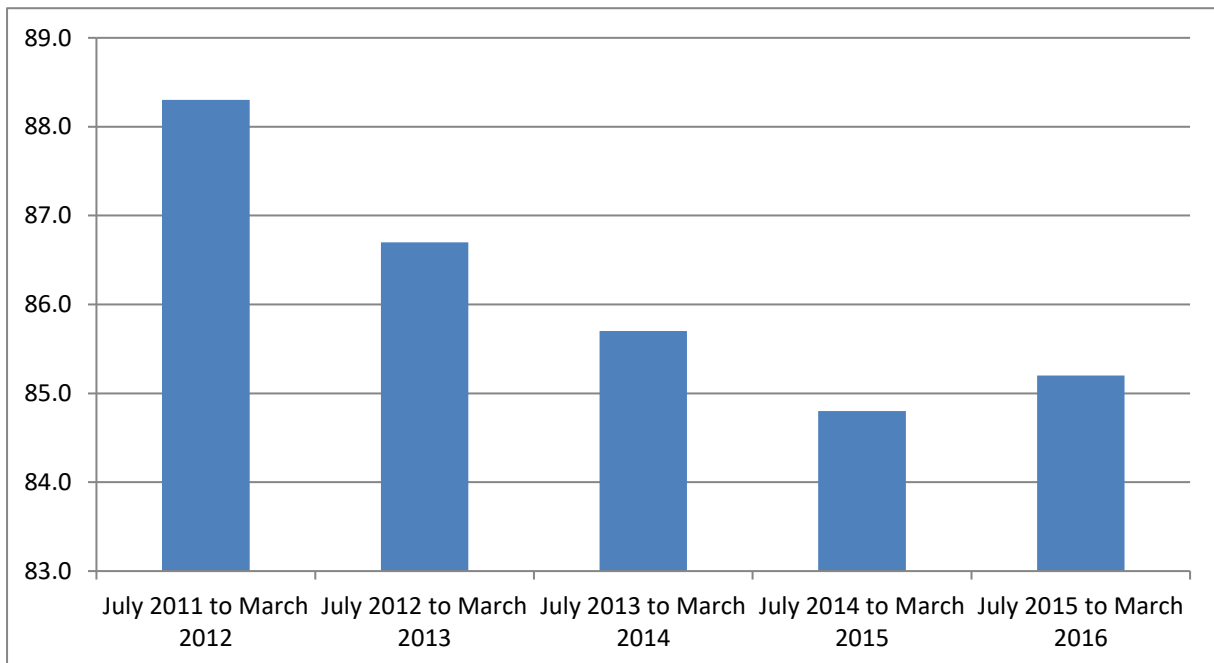

Figure A11: Weighted percentage of people reporting 'fairly good' or 'very good' overall experience of GP services

Comments
Small but significant decreases in each of the last three years, before picking up in the last year of the survey.

Domain 4 Overreaching Indicators	
Indicator:	Patient experience of dental services
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	GP Patient Survey (GPPS) from Ipsos MORI (http://www.gp-patient.co.uk) – Official Statistics
Years covered:	2011/12-2015/16 (financial year – period coverage July to September of previous year and January to March of following year for the respective financial year)
Recent changes:	N/A

Definition
This indicator measures the weighted percentage of people who report their overall experience of NHS dental services as 'fairly good' or 'very good'

Purpose
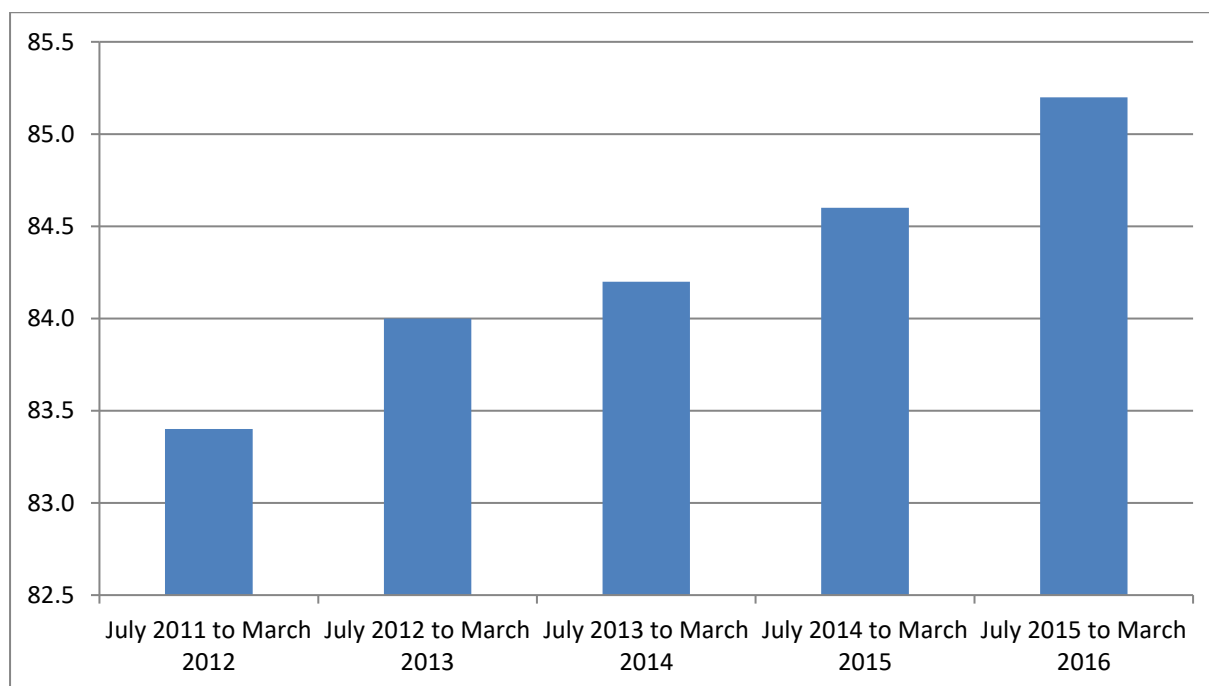This indicator aims to capture the experience of patients of NHS dental services.

Figure A12: Weighted percentage of people reporting 'fairly good' or 'very good' overall experience of NHS dental services

Comments
Time series shows small but significant increase of 2.2 percentage points from 2011/12 to 2015/16.

Domain 4	
Overreaching Indicators	
Indicator:	Patient experience of hospital care
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data Source:	National Inpatient Survey provided by the Care Quality Commission – Official Statistics. Published annually, available from Care Quality Commission (CQC) in April each year, although the 2014-15 survey was not released until May 2015 due to the general election.
Years covered:	2003/04-2015/16(financial year - Hospital stay 01/06 to 31/08; Survey collected 01/09 to 31/01 of the respective year)
Recent changes:	N/A

Definition
Patient experience measured by scoring the results of a selection of questions from the National Inpatient Survey looking at a range of elements of hospital care.

Purpose
This indicator aims to capture the experience of patients who have received medical treatment in hospital.
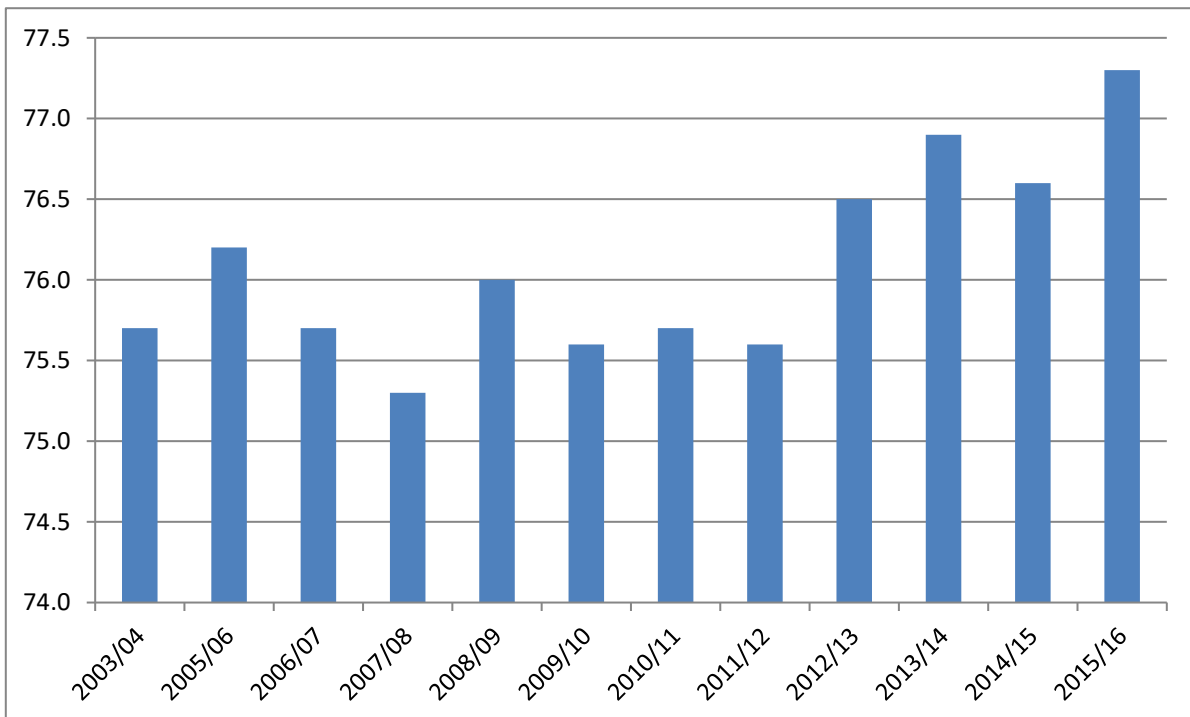

Figure A13: Average score of patients experience of hospital care

Comments
Value has fluctuated throughout the time series. Very little change from initial score in 2003/04.

Domain 5	
Improvement areas: Reducing the incidence of avoidable harm	
Indicator:	Deaths from venous thromboembolism (VTE) related events within 90 days post discharge from hospital
Source:	NHS Outcomes Framework (last accessed 8 Jun 2017)
Data Source:	Hospital Episode Statistics (HES) - National Statistics. Linked HES-ONS Mortality Data. Both Health and Social Care Information Centre (HSCIC).
Years covered:	2007/08-2015/16 (financial years)
Recent changes:	N/A

Definition

The indicator measures the number of patients who have been admitted to hospital with any cause and die within 90 days of their last discharge from a VTE related event, expressed as a rate per 100,000 adult hospital admissions.

Purpose

This indicator aims to measure the reduction in deaths from VTE related events by driving efforts to improve the prevention, detection and treatment of VTE before it causes death.
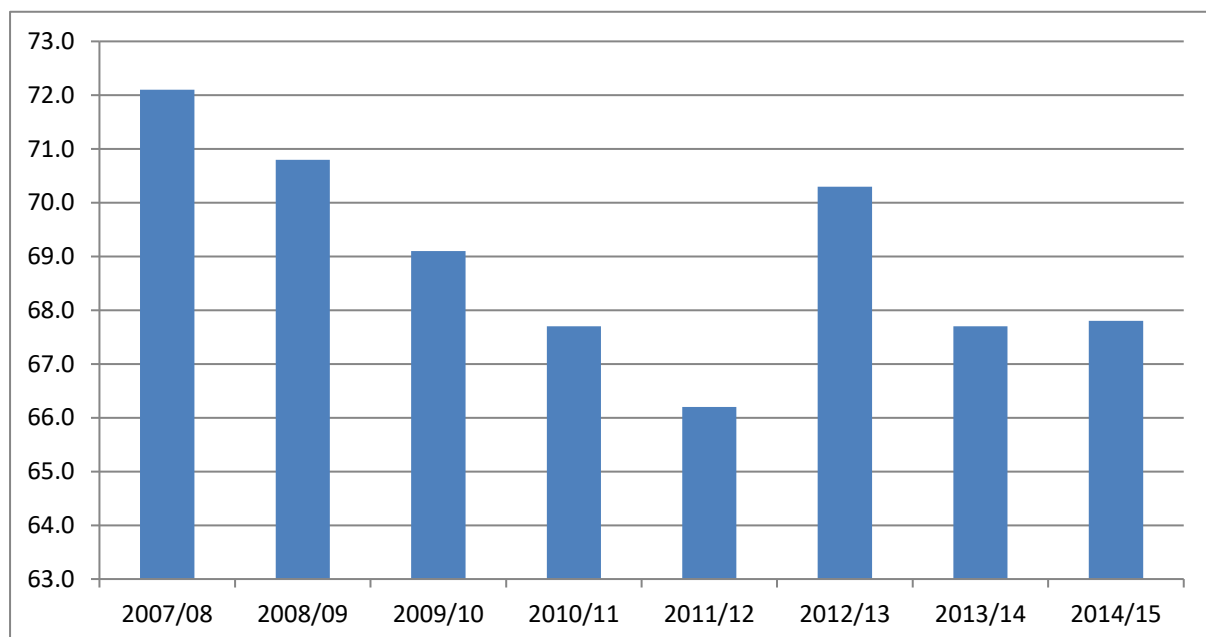

Figure A14: Deaths from VTE related events within 90 days post discharge. Crude rate per 100,000 (adult) hospital admissions

Comments

Indicator value decreasing between 2007/08 and 2011/12, with a spike in 2012/13 before settling at pre-2011/12 figures.

Domain 5	
Improvement areas: Reducing the incidence of avoidable harm	
Indicator:	Incidence of healthcare-associated infection - MRSA
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data Source:	Mandatory Surveillance of Healthcare Associated Infections by Public Health England (PHE)
Years covered:	2008/09-2015/16(Financial year)
Recent changes:	N/A

Definition

The number of Methicillin-resistant Staphylococcus aureus (MRSA) bloodstream infections reported to Public Health England (PHE).

Purpose

This indicator aims to measure the progress in reducing the incidence of avoidable harm, specifically healthcare-associated MRSA infections.
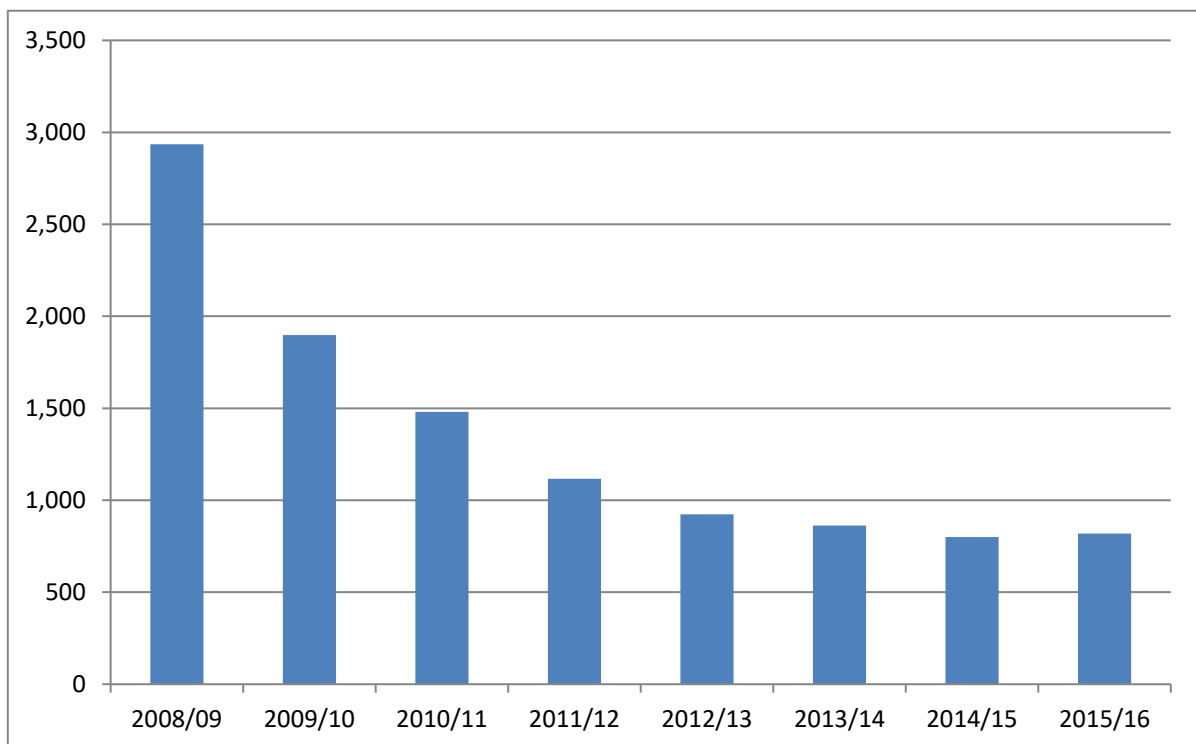

Figure A15: Crude count (Incidence) of all cases of MRSA blood stream infections reported

Comments

Large year on year decreases over the time series, a total reduction of 72percent from 2008/09. The pace of decrease is now slowing down.

Domain 5
Improvement areas: Reducing the incidence of avoidable harm

Indicator:	Incidence of healthcare-associated infection - C. difficile
Source:	NHS Outcomes Framework
	(last accessed 8 Jun 2017)
Data Source:	Mandatory Surveillance of Healthcare Associated Infections by Public Health England (PHE)
Years covered:	2007/08 – 2014/15 (financial year)
Recent changes:	N/A

Definition
The number of Clostridium difficile infections reported to Public Health England (PHE), in patients aged two years or older.

Purpose
This indicator aims to measure the progress in reducing the incidence of avoidable harm, specifically healthcare-associated C. difficile infections.
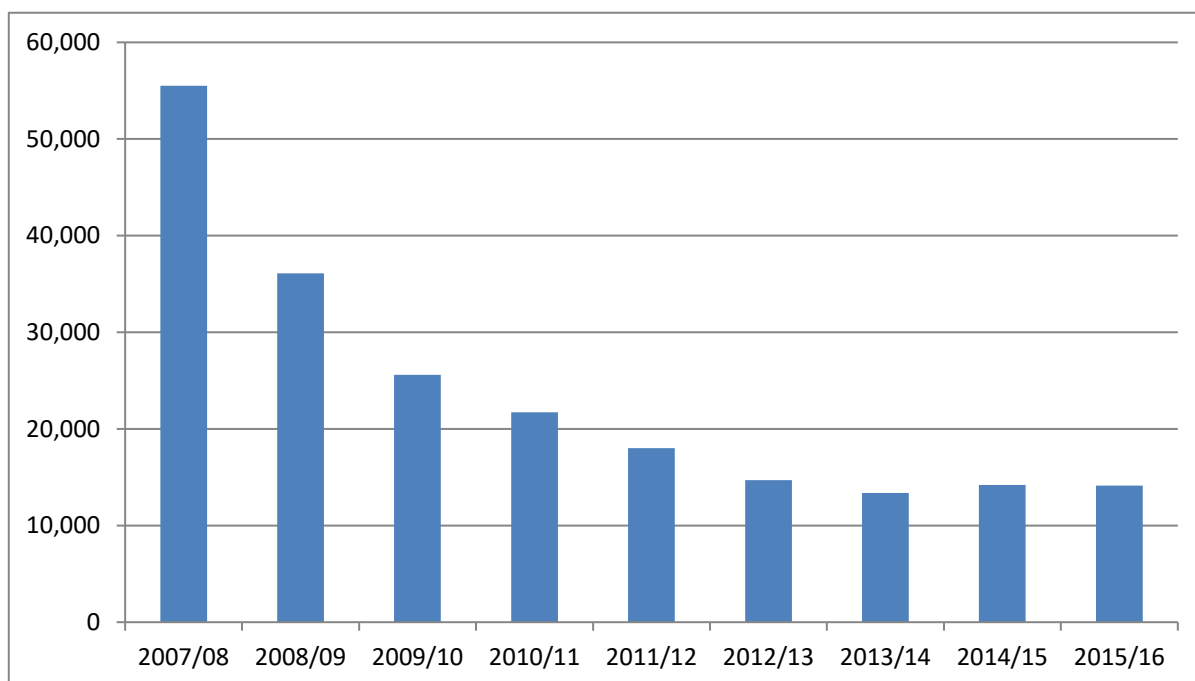

Figure A16: Crude count (incidence) of all cases of C-difficile infections reported

Comments
Large year on year decreases since 2007/08 (start of the time series) until 2013/14, since 2014/15 the number of C.difficile cases has started to increase.

Domain 5
Improvement areas: Improving the safety of maternity services

Indicator:	Admission of full-term babies to neonatal care
Source:	NHS Outcomes Framework
	(last accessed 8 June 2017)
Data Source:	The total number of full-term babies (gestation greater than 36 weeks) based on birth registration information linked to birth notifications data from the Office for National Statistics (ONS).
Years covered:	2010 – 2012 (calendar year)
Recent changes:	N/A

Definition
The number of full-term babies (gestation greater than 36 weeks) admitted within 28 days of birth to a neonatal unit, expressed as a percentage of all full-term births.

Purpose
This indicator aims to reduce the number of avoidable admissions of full-term babies to neonatal care and overall improve the safety of maternity services.
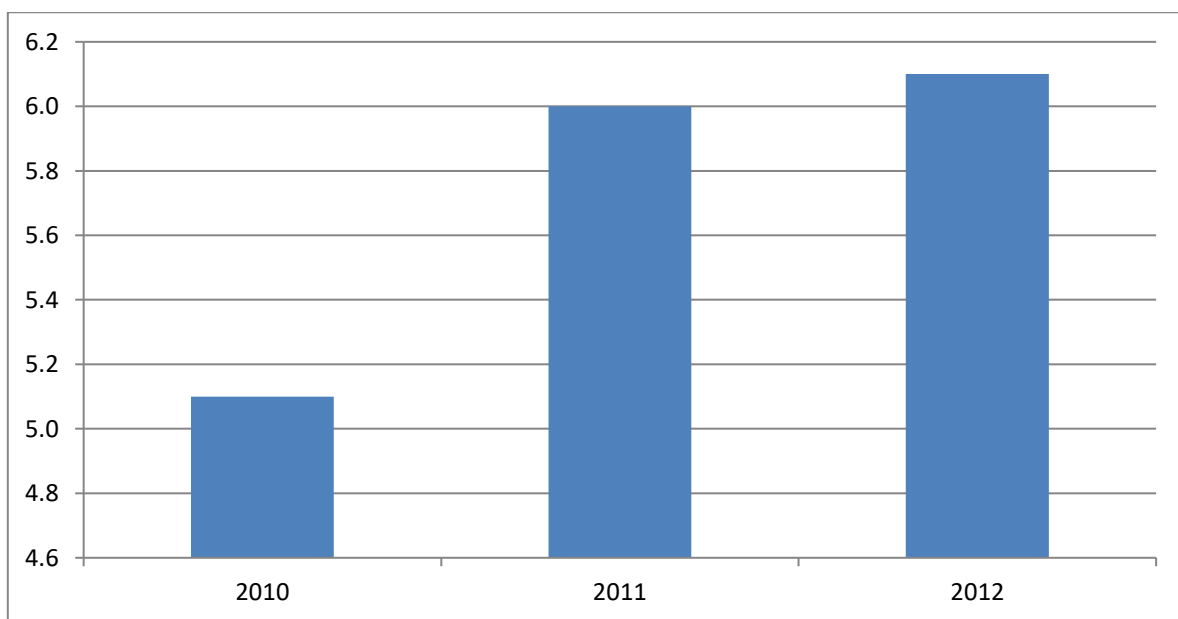

Figure A17: Admission of full-term babies to neonatal care. Full-term neonatal episodes rate per 100 full-term live births

Comments
Value appears to have increased slightly – interpret with caution, as the increase might be due to more admissions being reported rather than occurring.
Shortcoming: Only three data points are available.