Primary Health Care Research & Development 2014; **15:** 396–405
doi:10.1017/S146342361400005X

RESEARCH

Feasibility of applying review criteria for depression and osteoporosis national guidance in primary care

Mark F. Lambert[2], **Julia V. Cook**[1], **Ella Roelant**[1], **Colin Bradshaw**[4], **Robbie Foy**[3] and **Martin P. Eccles**[1]

[1]Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK
[2]Sunderland City Council, Sunderland, UK
[3]Leeds Institute of Health Sciences, University of Leeds, Leeds, UK
[4]Marsden Road Health Centre, South Shields, UK

Background: Data on the uptake of clinical guidelines into practice are essential to guide and evaluate quality improvement interventions. Organizations responsible for service specification, monitoring and improvement need to consider the practicality of and trade-offs made in different data collection methods. We examined the feasibility of deriving and applying review criteria for clinical guidelines in English primary care. **Methods:** We selected two sets of guidance, on osteoporosis and depression, and used a consensus process to derive review criteria. We manually extracted data on adherence to review criteria from patient records in 20 general practices from three NHS primary care trusts in northern England. We compared the relative utility of extracted data with that of routinely available data, summarizing feasibility using what we termed a Resource Ratio. **Results:** Of 53 proposed review criteria we assessed, 41 were judged clinically important, valid, relevant and measurable. Thirty-one could be assessed in 10% or more of sampled patients, whereas 15 could be readily extracted (resource ratio of 15 or less). Only eight met all desirable attributes for use as review criteria. Resource ratios correlated poorly with local stakeholders' prior views on feasibility of data collection. We observed wide variations in compliance with review criteria, with notably low levels among self-care standards. **Conclusions:** A minority of guideline recommendations were suitable for review criteria development, fewer still when using routinely available data. Local stakeholders tend to underestimate the actual resource requirements of data collection. Although improved design and use of clinical records may facilitate measurement of adherence to recommended practice, detailed assessments are still likely to rely upon some degree of manual data collection in the foreseeable future.

Key words: clinical guidelines; Primary Health Care; quality assurance; review criteria

Received 9 July 2013; revised 18 December 2013; accepted 5 January 2014; first published online 14 February 2014

Introduction

There are repeated policy calls for accelerated adoption of innovation and evidence across healthcare (Darzi, 2008). Rigorously developed clinical guidelines have a key role in reducing the gap between evidence and practice (Shekelle *et al.*, 1999). However, active approaches are often required to change clinical practice (Grimshaw *et al.*, 2004). Measuring quality of care has an important role to play in securing change. It is required to identify inappropriate variations in practice, and target improvement endeavours and monitor their impact. In the absence of such data, guideline implementation strategies are best guess rather than data-driven.

Correspondence to: Mark F Lambert, Sunderland City Council, c/o Loftus House, Colima Ave, Sunderland SR5 3XB, UK. Email: mark.lambert@doctors.org.uk

Such a criticism has been levelled at primary care in the United Kingdom (Audit Commission, 2008), where responsibility for implementation is shared between the National Institute for Health and Care Excellence (NICE) and local commissioners. The latter formerly comprised primary care trusts (PCTs), now replaced by clinical commissioning groups (CCGs) in recent National Health Service (NHS) reforms (Department of Health, 2012).

The introduction of the Quality and Outcomes Framework (QOF) in 2004 brought a step change in the availability of data on quality of primary care by promoting structured recording in electronic records (British Medical Association, 2010). Yet the utility of routine data from such schemes is potentially limited by incomplete coverage of health problems (Doran *et al.*, 2006). Early work on the impact of NICE guidance suggested that routine data are usually insufficient to assess compliance (Sheldon, 2004).

Review criteria are 'systematically developed statements relating to a single act of medical care that is so clearly defined that it is possible to say whether the element of care occurred or not retrospectively in order to assess the appropriateness of specific healthcare decisions, services and outcomes' (Campbell, 2002). Review criteria have been used to assess quality of primary care services (Hutchinson *et al.*, 2003). They can be developed from guideline recommendations and some NICE guidelines already suggest criteria which could, in principle, be readily adapted for quality measurement (Campbell, 2002). However, experience indicates that review criteria developed by expert panels may be 'unoperationalisable, unreliable, too rare to be useful, or too hard to extract reliably' (Campbell *et al.*, 2002). Indeed development work in this field identifies that much work is needed to operationalize proposed quality standards and that there are resource implications in collating and reporting review criteria (Rolfe, 2001).

We developed a set of review criteria to monitor the implementation of national guidance and investigated the feasibility of their application.

Methods

Selection of clinical topics

We selected two sets of guidance, on osteoporosis and depression, based on the availablility of national guidelines, population burden, relevance to primary care and likely potential for health gain.

Fractures caused by osteoporosis affect one in two women and one in five men over the age of 50 (Cummings and Melton, 2002) costing the NHS £1.8 billion annually (Dolan and Torgerson, 1998). Primary care is mostly responsible for on-going clinical management but there is considerable potential scope for improving reliability of primary care case-finding and primary and secondary prevention (National Institute for Health and Clinical Excellence, 2011). Depression is ranked as the fourth leading cause of burden among all diseases and is expected to show a rising trend during the coming 20 years. Estimates of prevalence range from 2% for major depression to 10% for mixed depression and anxiety. NICE guidance highlights a number of key priorities for implementation, improving diagnosis, drug treatment and self-help (National Institute for Health and Clinical Excellence, 2009).

Development of review criteria

Based on the clinical guidance recommendations, we produced a list of candidate review criteria from the relevant guidelines: 43 covering osteoporosis (Compston *et al.*, 2010); and 71 covering depression (National Institute for Health and Clinical Excellence, 2009).

We convened a consensus panel for each condition comprising specialist clinicians, primary and community health service clinicians (including GPs) and managers, and a patient representative. We used a modified RAND consensus process (Murphy *et al.*, 1998). Each set of panellists initially independently rated candidate criteria as a postal exercise. The criteria were measured on three parameters: clinical importance; importance of recording; and ease of measurement. A scale of 1–9 was used for rating each criterion characteristic (eg, 1 = low importance, 9 = high importance for the criterion of clinical importance).

Candidate criteria scoring seven or more for both clinical importance and importance of recording and with panel consensus (not more than two outliers scoring less than seven) were taken forward without further discussion into the data collection phase.

Where consensus was insufficient on one or more ratings (more than two panellists rating

https://doi.org/10.1017/S146342361400005X Published online by Cambridge University Press

outside the three point range of the median score 1–3, 4–6 and 7–9), criteria were independently re-rated during a 3-hour face-to-face structured panel meeting. Where additional review criteria were suggested by the consensus panel (five for osteoporosis and four for depression), the panel debated whether to rate these as well.

Candidate criteria were dropped if case note review was not considered feasible by the research team (eg, 'practitioners build a trusting relationship and worked in an open, engaging and non-judgmental manner'.)

Criteria with high median final consensus scores for both clinical importance and importance of recording (scoring seven or higher) were taken forward for data collection. Before doing so, the research team (J.C. and M.P.E.) rated each review criterion for clinical importance (by subtracting six from the consensus score, with '2' being the minimum rating for inclusion) and for relevance to the primary care sector with those relating to secondary care rated 1 (below the minimum threshold for inclusion), those relating to the interface between primary and secondary care were rated '2' and those relating to primary care were rated '3.'

Data collection

We recruited general practices from Gateshead, South Tyneside and Sunderland in the North of England through their involvement in the consensus process, the Primary Care Research Network, and publicity at local continuing medical education events.

Practices identified potential participants with osteoporosis from a pre-defined computer search. This search sought those aged over 55 years with a relevant Read Code in the preceding five years (osteoporosis, fragility fracture or fracture of hip, spine, pelvis or arm). Each practice posted invitations to a one in four sample of patients, aiming for a total practice sample of 50.

We took a similar approach to identify patients with depression. We identified patients aged over 18 years with a QOF Read Code for depression or commonly used depression codes during the preceding 12 months. Practices posted 70 invitations to systematically identified patients with depression. Two practices with fewer than six consenting patients per condition sent further batches of invitations.

The review criteria were translated into measurable data items. These were extracted from individual patient records using a structured data collection form by a single, medically qualified data collector (J.C.). Descriptive summaries of the data were compiled using Microsoft Excel, and all statistical analyses were undertaken with Stata 9.2.

Practice level data on QOF performance were available from the Health and Social Care Information Centre (Information Centre for Health and Social Care, 2012). Practice level prescribing data on volume and costs of classes of drugs and individual agents were available from the Electronic Prescribing Analysis and Cost (ePACT) system (NHS Business Services Authority, 2008) for the 20 participating general practices. We focused on bone sparing agents and antidepressants (excluding amitriptyline as this is used for other therapeutic fields). We calculated prescribing rates based upon practice size.

Assessing feasibility, validity and frequency of measured review criteria

We assessed three aspects of feasibility: complexity, ease of locating data and method of data collection. First, we examined how many data items were required to establish criterion-compliance; the greater the number of items, the greater the complexity. For example, the formal risk assessment or DEXA scanning of people with rheumatoid disease (OI 7) is a complex review criterion (scoring '1') in that it includes an age (over 65) and diagnostic (rheumatoid arthritis) criteria defining the eligible cohort, as well as two options for fulfilling the review criterion (a DEXA scan or an outpatient risk assessment) and a time limit for doing so (in the last three years). More straightforward criteria scored '2.'

Second, we established the ease of locating data in medical records. Where information was required from multiple locations or where it was inconsistently recorded, we assigned the lowest score ('1'). For example, to establish compliance with OI5 (blood tests for osteoporosis) we had to check computerized laboratory results, continuation notes and letters from secondary care. In contrast, review criteria for which consistently recorded data could be extracted from a single location scored highest at '3'.

https://doi.org/10.1017/S146342361400005X Published online by Cambridge University Press

Third, we assessed the method of data collection, where we needed to undertake manual data extraction from non-coded material or free text data entries, the lowest scoring ('1') was given. Criteria only requiring data from simple electronic extraction (eg, routinely identified single Read codes) scored '3'.

We pragmatically summarized these three considerations by producing an aggregate score for each criterion. In the absence of an alternative model, we generated what we have termed a 'resource ratio' from the sum of the three descriptors of feasibility (number of data items, location of data and ease of extraction). Each descriptor was rated between '1' and '3' (between '1' and '2' for number of components) so that a maximum score of 8 represented the least resource intensive review criteria. As the primary consideration was resource use, we converted this to a resource ratio by taking the reciprocal of this sum and multiplying by 100 for convenience. In summary:

$$\text{Resource ratio} = \frac{100}{(\text{component rating} + \text{location rating} + \text{extraction rating})}$$

We compared this with the panellists' prior views of the ease of obtaining data by a scatter plot of the resource ratio against their scores (Figure 1).

We rated validity of criteria based upon the degree of extrapolation required for interpretation. Where we had to make significant extrapolation from the available data to match the content of the criterion (eg, use of antidepressants in specific subgroups) a validity rating of '1' was given (and this rating was considered below the minimum threshold for inclusion). For example, presentation with symptoms of longer than two years (RC13) was assessed by reviewing those patients with mild depression of the chronic continuous subtype. Where the data we collected matched most of the content of the criterion, with little need for extrapolation, we assigned a validity rating of '2.' And where the data we obtained exactly matched all of the content of the criterion, for example, use of generic SSRIs first line (RC16) we assigned a rating of '3'.

Finally, we highlighted which criteria were relevant to <10% of cases and hence less likely to be relevant for wider use.

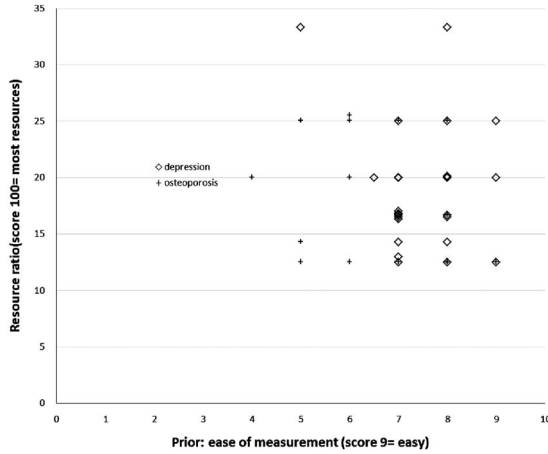

Figure 1 Ease of obtaining data: prior views versus resources used.

Additional interpretation was required to classify patients with depression, as many review criteria require the nature of the condition to be defined. Therefore, before assessing compliance with depression review criteria, patients were classified by their status at the time of the case note review: according to whether they appeared as new episode of depression (ie, either a first presentation or previous single episode of depression over five years ago), had continuous episodic depression (ie, patients with one or more discrete episodes of depression in the last five years) or had continuous chronic depression (ie, patients with one or more episodes of depression with continuous symptoms).

Results

Characteristics of participating practices and patients

Of the 20 participating practices, patient list sizes and staffing complements were slightly higher than the national means and training practices were over-represented (Table 1).

There was a six-fold variation between the 20 practices in coded prevalence of both osteoporosis and depression. From these practices, 160 patients were recruited for each condition. Tables 1 and 2 show the key characteristics of practices and participating patients, respectively. Tables 3 and 4

https://doi.org/10.1017/S146342361400005X Published online by Cambridge University Press

Table 1 Practice and patient characteristics

	Sample (n = 20)	Comparison
List size: mean (95% CI)	7171 (5292, 9050)	6651*
Whole time equivalent GPs: mean (95% CI)	4.6 (3.3, 5.9)	4.3*
Whole time equivalent nurses: mean (95% CI)	2.2 (1.7, 2.8)	1.6*
Training practices	12 (60%)	36 (33%)**

Table 2 Patient characteristics

Osteoporosis		
Practice prevalence (per 1000 on list): median (range)	19 (5, 31)	
Practice level response rates to participation	5–58%	
Patients recruited by practice: mean (95% CI)	8 (6.7, 9.3)	
Gender	Male 14%	Female 86%
Age – mean	68 years	
Presence of fracture	83%	
Depression		
Practice prevalence (per 1000 on list): median (range)	50 (13, 77)	
Practice level response rates to participation	6–28%	
Patients recruited by practice: mean (95% CI)	7.5 (6.4, 8.6)	
Gender	Male 30%	Female 70%
Age – mean (95% CI)	52 years (49.9, 54.6)	
Classification of depression at time of case note review		
New episode of depression	30%	
continuous episodic depression	43%	
continuous chronic depression	27%	

*2011 figures from General and Personal Medical Services England 2001–2011.
**Figures for Gateshead, South Tyneside and Sunderland primary catre trusts supplied by Northern Deanery.

summarize the ratings for the review criteria, their resource ratios and measurements of adherence.

Osteoporosis review criteria

Twenty-three out of 28 candidate criteria were judged both clinically important and important to record. Of these 23 proposed standards (Table 3), 18 had data from one or more patients in the sample. Sixteen of these were judged clinically important by the review panels and a valid representation of clinical practice, of which a further 15 were also relevant to primary care or the interface of care. Ten of these could be assessed in 10% or more of the sample.

Eight of the 18 fell in to the most feasible group in each of the three categories (complexity, location and extraction). Four of these (two relating to drug treatments and two to organization of care) with a resource ratio of 15 or less were in the cluster of indicators that were most valid, relevant and could be applied to 10% or more of cases reviewed.

Depression review criteria

Thirty out of 74 candidate criteria were judged both clinically important and important to record. Of 30 initially proposed criteria, data were available on 29 (Table 4). Twenty-five of these were considered to be a valid measure of clinical practice. Twenty-four were relevant either to primary care or the interface of care. Twenty-one could be assessed in 10% or more of the sample.

Four indicators fell in to the most feasible group in each of the three categories (complexity, location and extraction) and had a resource ratio of 15 or less. All four of these (two relating to drug treatments and two to organization of care) were in the cluster of indicators that were most valid, relevant and could be applied to 10% or more of cases reviewed.

https://doi.org/10.1017/S146342361400005X Published online by Cambridge University Press

Table 3 Field testing osteoporosis review criteria

Identity number	Indicator name	*n*	*N*	Criterion met	Feasibility			Prior views: Ease	Rating				Resource ratio
					Components	Locations	Extraction		Importance	+ Validity	+ Sector	+ Frequency	
OR17	**Alendronate prescribed**	62	67	93%	2	3	3	9	+	+	+	+	12.5
OD2	**Fracture register**	17	50	34%	2	3	3	5	+	+	+	+	12.5
OR15	**Ca/Vit D prescribed**	91	132	69%	2	3	3	8	+	+	+	+	12.5
OD21	**Medication review**	145	160	91%	2	3	3	7	+	+	+	+	12.5
OS 1	Evidence-based patient info	22	160	14%	2	2	1	4	+	+	+	+	20.0
OS12	Weight bearing exercise advice	12	160	8%	2	1	1	5	+	+	+	+	25.0
OA 6	Family history of hip fracture	6	160	4%	2	1	1	5	+	+	+	+	25.0
OD22	Medication follow-up	12	128	9%	1	2	1	6	+	+	+	+	25.0
OI 10	DEXA scan for high risk	19	22	86%	1	1	2	6	+	+	+	+	25.0
OA 14	Falls assessment in osteoporosis	18	41	44%	1	1	2	8	+	+	+	+	25.0
OR 18	Bisphosphonate for alendronate intolerant	9	13	69%	2	3	3	8	+	+	+	0	12.5
OR 19	Bisphosphonate intolerance where raloxifence or strontium prescribed	6	8	75%	2	3	3	8	+	+	+	0	12.5
OS 13	Advice on excess alcohol intake	2	3	67%	2	3	3	6	+	+	+	0	12.5
OR 20	Bone sparing if on steroids	7	7	100%	1	3	2	8	+	+	+	0	16.7
OA 8	Osteoporosis assessment if on steroids	6	8	75%	2	1	2	6	+	+	+	0	20.0
OI 7	DEXA or formal risk assessment in rheumatoid	4	5	80%	1	1	2	7	+	+	0	0	25.0
OA 9	**BMI measured**	115	160	72%	2	3	3	8	0	0	0	0	12.5
OA 11	FRAX if postmenopausal or over 50	13	160	8%	2	2	3	5	0	0	0	0	14.3

Bold values represent the review criteria that meet all desirable attributes.

Table 4 Field testing depression criteria

Identity number	Indicator name	*n*	*N*	Criterion met	Feasibility			Prior views:	Rating				Resource ratio
					Components	Locations	Extraction	ease	Importance	+ Validity	+ Sector	+ Frequency	
DR16	**Generic SSRIs first line**	116	143	81%	2	3	3	8	+	+	+	+	12.5
DD 6	**Case identification from disease registers**	24	28	86%	2	3	3	7	+	+	+	+	12.5
DR20	**Different SSRI or better tolerated new generation**	27	37	73%	2	3	3	7	+	+	+	+	12.5
DO31	**Continuity of care**	135	150	90%	2	3	2	7	+	+	+	+	14.3
DA 3	Suicide risk assessment	45	150	30%	2	2	2	7	+	+	+	+	16.7
DO10	Appropriate quantity of prescription	2	29	7%	2	2	2	8	+	+	+	+	16.7
DR15	Talking and pharmacological therapy (moderate/severe cases)	66	93	71%	2	2	2	7	+	+	+	+	16.7
DR24	Talking + pharmacotherapy where one failed alone	101	150	67%	2	2	2	7	+	+	+	+	16.7
DS 9a	Antidepressant side effects explained	31	86	36%	2	3	1	7	+	+	+	+	16.7
DS9b	Antidepressant side effects of suicidal ideation explained	2	86	2%	2	3	1	7	+	+	+	+	16.7
DS17a	Advised on potential side effects	31	86	36%	2	3	1	7	+	+	+	+	16.7
DS17b	Advised on potential interactions	6	86	7%	2	3	1	7	+	+	+	+	16.7
DO 26	Continuing therapy six months after relapse	26	143	18%	2	3	1	8	+	+	+	+	16.7
DO 18	Review appointments arranged	41	109	38%	1	2	2	8	+	+	+	+	20.0
DS 4	Safety net arrangements in case of suicidal ideation	14	150	9%	2	2	1	7	+	+	+	+	20.0
DR 21	Appropriate switching between antidepressants	14	54	26%	2	2	1	6.5	+	+	+	+	20.0
DO 27	Review of antidepressants, six months after remission	5	31	16%	2	2	1	8	+	+	+	+	20.0
DO 25	Specialist referral for most severe/complex cases	19	39	49%	1	2	2	9	+	+	+	+	20.0
DO 5	Contact following DNA	12	40	30%	2	1	1	8	+	+	+	+	25.0
DO 19	Check and step up dose in initial non-response	5	172	3%	1	2	1	7	+	+	+	+	25.0
DA 1 +2	Comprehensive assessment	36	150	24%	1	1	1	5	+	+	+	+	33.3
DO 8	Urgent referral in high risk	11	14	79%	2	3	3	9	+	+	+	0	12.5
DR 12	No antidepressants in recent onset mild cases	2	10	20%	2	3	2	8	+	+	+	0	14.3
DO 11	Additional support in suicidal ideation	6	14	43%	2	2	1	8	+	+	+	0	20.0
DO 22	Treatment augmentation in consultation with specialist	9	150	6%	2	1	1	9	+	+	0	0	25.0
DO 14	High-intensity psychol or pharmacotherapy in initial psychol failure	30	57	53%	1	2	2	7	+	0	0	0	20.0
DO 28	Two-year treatment in high risk of relapse	19	39	49%	2	2	1	7	+	0	0	0	20.0
DO 7	Mental health assessment were screen positive	5	6	83%	2	2	1	7	+	0	0	0	20.0
DR 13	Antidepressants were specifically indicated	7	8	88%	1	1	1	8	+	0	0	0	33.3

Bold values represent the review criteria that meet all desirable attributes.
Identity number = code assigned following consensus process (links to indicator name in supplementary file).
n = number of cases meeting this criterion.
N = number of records in which criterion is reported or applicable.
Components = number of components to required to assess criterion (rating).
Location = locations required to assess criterion (rating).
Extraction = method of data extraction (rating).
Prior views: ease = panel rating of ease of data collection before data extraction.
Importance = clinical importance rating (meets threshold + or not 0).
+ Validity = validity assessment rating (meets threshold + or not 0).
+ Sector = sector that the criterion applies to (meets threshold + or not 0).
+ Frequency = frequency of reporting (meets threshold + or not 0).

Study resources

Some of the resources required for this research study would not be routinely required to undertake assessment of review criteria in a service setting. These include time and effort in developing the review criteria, such as the extraction of candidate criteria from clinical guidance and the conduct and analysis of the consensus panels.

We found no clear relationship between the resource ratio and the panels' prior assumptions about ease of accessing data and the level or resources documented either for depression or for osteoporosis (Figure 1).

Contribution of routinely available data

Routine prescribing data did not assist in establishing performance against any of the identified review criteria. Data published on the depression standards in the clinical domain of QOF provided an estimate of the extent of case finding. Returns made by the 20 practices in this sample for 2009/10 QOF payments declared 91.7% achievement on the standard relating to prompt assessment, with 3.3% of patients excluded, giving an overall 88.6% compliance with this review criterion. But neither the use of severity measures and repeat measures of severity used in QOF (British Medical Association, 2013) appeared in the review criteria proposed by the panel.

Discussion

We were able to assess clinical quality for two sets of national guidance, irrespective of pre-existing routine data coverage. We developed review criteria that were valid representations of primary care quality, and demonstrated which were readily available from clinical records. However, only four of a total of 23 review criteria for osteoporosis and only four of 30 review criteria for depression met all of our requirements. We had anticipated that local expert opinion would be useful in predicting the level of resource used in obtaining data but found otherwise. However, we have identified issues that should be actively considered in planning to use review criteria.

This study primarily sought to examine feasibility of developing and applying review criteria. Although we have identified variations in standards of care, the sampling of practices and patients is not necessarily representative of prevailing standards of care. Comparison with routinely available data was possible, albeit on only one topic indicator.

We have built on existing literature on developing and using review criteria (Rolfe, 2001; Campbell *et al.*, 2002; Hutchinson *et al.*, 2003) and have gone further in demonstrating a practical way to assess and summarize feasibility in routine practice, using a resource ratio. Further, we have illustrated the difficulties in applying review criteria to assessing standards of practice in relation to national guidance.

Our study had four main limitations. First, we only examined practice relating to two guidelines in one locality, thereby limiting the generalizability of our findings. Second, there is a risk of selection bias given that participating practices were more likely than average to be larger, better staffed and involved in training and that we could only collect data for consenting patients. Therefore, our findings probably overstate the quality or availability of information about quality of care. This is unlikely to invalidate our analysis of the feasibility and validity of applying the criteria. If similar methods were used as part of a clinical audit programme in the United Kingdom and elsewhere (rather than the research format used here), there would be neither the need to seek formal patient consent nor any need for the resources to do so. Third, not everything that is important can be measured, especially humanistic or holistic aspects of care. However, we went through a formal and rigorous process involving clinicians to prioritise those criteria that they judged were clinically important and meaningful. Despite having done this, it was still not possible to collect all of the data. Fourth, if this approach were used to determine quality in primary care, other confounding factors should be taken into account. For example, many aspects of good clinical care for people with depression or osteoporosis are dependent on referral for diagnosis or specialist care. Therefore, compliance against review criteria could be significantly influenced by the availability or perceived quality of local diagnostic or specialist services.

The methods carried out here should be representative of those required for analysis of a quality of care using an electronic primary care record. We selected common conditions, which we believe to be representative of routine record keeping practice.

https://doi.org/10.1017/S146342361400005X Published online by Cambridge University Press

NICE has been given the task of developing Quality Standards to monitor the quality of care in the United Kingdom. Only four out of the 13 proposed quality standards for depression in adults (National Institute for Health and Clinical Excellence, 2013a) satisfy requirements for our review criteria and three (Quality Standards 6, 10 and 11) meet our threshold resource ratio of 12 or less. Seven of the remaining Standards were not included in our initial list of candidate criteria because of significant problems, such as insufficient clarity about the target population.

NICE uses different methods to those in this study for developing its Quality Standards (National Institute for Health and Clinical Excellence, 2013b) but still the discrepancies between our review criteria and the Quality Standards are significant. Equally concerning, is the observation that the standards developed by NICE are likely to be difficult to operationalize because of the population definitions used. This finding was unexpected, as in other aspects of its work NICE has undertaken field testing before using measurements of primary care quality (Sutcliffe *et al.*, 2012).

This UK-based study suggests those responsible for primary care quality should not take publication in national standards as an indication that the data required to establish performance against these standards are available. Views of local experts on ease of access to information about quality of care should also be used with caution. All quality standards should be assessed in routine practice for the complexity, ease and method of data extraction before proposing them as substantive review of quality.

Implications for research

There is an increasing interest in routinely collected clinical data in primary care for both service monitoring and research. Even allowing for improvements in the sophistication of data entry and collection, efforts to undertake comprehensive assessments of the quality of care are still likely to rely upon manual data collection to some degree in the foreseeable future.

Our methods for rating feasibility of data collection (resource ratio) and assessing their validity require further evaluation in different sets of guidance and more representative samples of patients and practices.

Conclusions

Assessing feasibility and resource use is a neglected step in the development of quality standards and review criteria. They are difficult to predict and require piloting as an essential step in the development of what should otherwise be useful tools for quality improvement.

Acknowledgements

The authors thank the practices who participated and the patients, who gave permission for access to their records, and Lucy Topping and Janette Stephenson, who contributed to the development of this project.

Funding

This work was supported by a grant from NHS South of Tyne and Wear.

References

Audit Commission. 2008: *Is the treatment working? Progress with the NHS reform system*. London: Audit Commission.

British Medical Association. 2010: *General Practitioners – briefing paper*. London: British Medical Association.

British Medical Association. 2013: QOF guidance previous revisions [WWW Document]. Retrieved 17 May 2013 from http://bma.org.uk/practical-support-at-work/contracts/independent-contractors/qof-guidance/qof-guidance-previous-revisions

Campbell, S.M. 2002: Research methods used in developing and applying quality indicators in primary care. *Quality and Safety in Health Care* 11, 358–64.

Campbell, S.M., Hann, M., Hacker, J., Durie, A., Thapar, A. and **Roland, M.** 2002: Quality assessment for three common conditions in primary care: validity and reliability of review criteria developed by expert panels for angina, asthma and type 2 diabetes. *Qualilty and Safety in Health Care* 11, 125–30.

Compston, J., Cooper, A., Francis, R., Kanis, J., Marsh, D., McCloskey, E., Reid, D., Selby, P. and **Wilkins, M.** 2010: Guideline for the diagnosis and management of osteoporosis in postmenopausal women and men from the age of 50 years in the UK. *National Osteoporosis Guideline Group (NOGG)*.

Cummings, S.R. and **Melton, L.J.** 2002: Epidemiology and outcomes of osteoporotic fractures. *The Lancet* 359, 1761–767.

Darzi, A. 2008. *High quality care for all: NHS Next Stage Review final report (No. Cm7432)*. London: Department of Health.

Department of Health. 2012: Functions of clinical commissioning groups [WWW Document]. Retrieved 26 February 2013 from http://www.dh.gov.uk/health/2012/06/ccg-functions/

Dolan, P. and **Torgerson, D.J.** 1998: The cost of treating osteoporotic fractures in the United Kingdom female population. *Osteoporos International* 8, 611–17.

Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U. and **Roland, M.** 2006: Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine* 355, 375–84.

Grimshaw, J.M., Thomas, R.E., MacLennan, G., Fraser, C., Ramsay, C.R., Vale, L., Whitty, P., Eccles, M.P., Matowe, L., Shirran, L., Wensing, M., Dijkstra, R. and **Donaldson, C.** 2004: Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technology Assessment* 8 (iii–iv), 1–72.

Hutchinson, A., McIntosh, A., Anderson, J., Gilbert, C. and **Field, R.** 2003: Developing primary care review criteria from evidence-based guidelines: coronary heart disease as a model. *British Journal of General Practice* 53, 690–96.

Information Centre for Health and Social Care. 2012: The Quality and Outcomes Framework 2010/11 [WWW Document]. Information Centre for Health and Social Care. Retrieved 21 August 2012 from http://www.ic.nhs.uk/qof

Murphy, M., Black, N., Lampling, D., McKee, C.M., Sanderson, C., Askham, J. and **Marteau, T.** 1998: Consensus development methods, and their use in clinical guideline development: a review. *Health Technology Assessment* 2(3).

National Institute for Health and Clinical Excellence. 2009: Depression in adults (update) [WWW Document]. NICE. Retrieved 3 December 2012 from http://www.nice.org.uk/

National Institute for Health and Clinical Excellence. 2011: Osteoporosis - secondary prevention including strontium ranelate [WWW Document]. NICE. Retrieved 3 December 2012 from http://www.nice.org.uk/

National Institute for Health and Clinical Excellence. 2013a: Depression in adults quality standard Introduction and overview QS8 [WWW Document]. Retrieved 31 January 2013 from http://publications.nice.org.uk/depression-in-adults-quality-standard-qs8

National Institute for Health and Clinical Excellence. 2013b: More information about the NICE quality standards [WWW Document]. Retrieved 4 July 2013 from http://www.nice.org.uk/guidance/qualitystandards/moreinfoaboutnicequalitystandards.jsp

NHS Business Services Authority. 2008: Prescription Services ePACT [WWW Document]. Retrieved 20 November 2013 from http://www.nhsbsa.nhs.uk/815.aspx

Rolfe, M.K. 2001: The NEBPINY programme (Phase 1). Changing practice – testing a facilitated evidence-based programme in primary care. *Journal of Clinical Governance* 9, 21–26.

Shekelle, P.G., Woolf, S.H., Eccles, M. and **Grimshaw, J.** 1999: Clinical guidelines: developing guidelines. *BMJ* 318, 593–96.

Sheldon, T.A. 2004: What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. *BMJ* 329, 999.

Sutcliffe, D., Lester, H., Hutton, J. and **Stokes, T.** 2012: NICE and the Quality and Outcomes Framework (QOF) 2009–2011. *Quality in Primary Care* 20, 47–55.