Journal of Evaluation in Clinical Practice, **5**, 3, 269–281

Achievable benchmarks of care: the ABC™s of benchmarking

Norman W. Weissman PhD,[1,2] Jeroan J. Allison MD, MS,[2] Catarina I. Kiefe PhD, MD,[2] Robert M. Farmer,[3] Michael T. Weaver PhD, MSN,[4] O. Dale Williams MPH, PhD,[2] Ian G. Child MSHA,[1] Judy H. Pemberton PhD,[4] Kathleen C. Brown PhD[4] and C. Suzanne Baker MPH[2]

Center for Outcomes and Effectiveness Research and Education (COERE), University of Alabama at Birmingham, Birmingham, Alabama, USA

[1]Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, Alabama, USA

[2]Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA

[3]Alabama Quality Assurance Foundation, Inc., Birmingham, Alabama, USA

[4]School of Nursing, University of Alabama at Birmingham, Birmingham, Alabama, USA

Correspondence
Dr Norman W. Weissman
University of Alabama at Birmingham
Health Service Administration
1675 University Boulevard
Birmingham, AL 35294-3361, USA

Keywords: benchmarks, CQI, performance measures, process of care, quality of care, quality improvement

Accepted for publication:
13 April 1999

Abstract

Benchmarking is generally considered to be an important tool for quality improvement. Traditional approaches to benchmarking have relied on subjective identification of 'leaders in the field'. We derive an objective, reproducible and attainable Achievable Benchmark of Care (ABC™) by measuring and analysing performance on process-of-care indicators. Three characteristics of the ABC™ that we deem essential are: (1) benchmarks represent a measurable level of excellence; (2) benchmarks are demonstrably attainable; (3) benchmarks are derived from data in an objective, reproducible and predetermined fashion. From these characteristics it follows that (4) providers with high performance are selected to define a level of excellence in a predetermined fashion, but (5) providers with high performance on small numbers of cases do not influence unduly benchmark levels. We use the 'pared mean' to operationalize the ABC™. Roughly, the pared mean summarizes the performance of top-ranked providers whereby at least 10% of the patient pool across all providers is included. Bayesian estimators for adjustment of performance of providers with small sample sizes are used to rank providers. Randomized controlled trials to assess the independent effect of the ABC™ in quality improvement projects are under way. We have developed a methodology objectively and reproducibly to derive a level of excellent, attainable performance, based on measured performance by a group of providers. The ABC™ can be applied to groups of providers in communities, to institutions and departments within them, or to individual practitioners.

Introduction

The Achievable Benchmark of Care (ABC™) method for expanding and improving the measurement of health care quality, is being refined under an initiative of the Agency for Health Care Policy and Research (AHCPR) (Clancy 1997). ABC™ enhances the modern health care quality improvement repertoire as a tool for translating evidence into medical practice. (Kiefe *et al.* 1998). The method can be used by providers, insurers or government agencies in a wide range of settings, including hospitals, physician practices, nursing homes, public health clinics or managed care organizations. ABC™s can

© 1999 Blackwell Science

be applied to groups of providers in communities, to institutions and departments within them or to individual practitioners. The ABC™ approach is objective, readily understandable, easily updated and useful in identifying performance areas amenable to improvement. It has been tested in several clinical conditions, with providers ranging from those included in a national hospital census to physicians in a single HMO.

In this paper, we provide a brief history of benchmarking, explain how the ABC™ builds on traditional concepts of benchmarking, discuss in detail the methodology for deriving ABC™s and, finally, give some practical application techniques with examples.

Traditional approach to benchmarking and quality improvement

Today, the health care sector is charged with general directives such as: 'Improve quality and outcomes of care'. It follows that providers, administrators and controlling agencies will ask: 'Improve in relation to what?' From this it is a short step to: 'What are realistic, achievable goals? Who is achieving or exceeding target quality now? How are they doing it? What are the standards of quality, and how do we set realistic improvement mile-posts along the way to improvement?'

Benchmarking, which directly addresses these issues and plays a central role in quality improvement (Mohr *et al*. 1996), has been defined as the identification of 'industry leaders' so that the leaders' practices may be understood and emulated (Berkey 1994). The concept of 'industry leaders' is subjective, and traditional Quality Improvement (QI) tools do not include a uniform data-driven definition of benchmark performance. Until recently, the available literature on benchmarking had not gone beyond definitions such as 'finding and implementing best practices' (Camp & Tweet 1994) or 'benchmarking is the continuous process of measuring products, services and practices against the toughest competitors or those known as leaders in the field' (Camp 1994). In fact, in many discussions of benchmarking, it is not clear whether process, outcome or both are being benchmarked. While hospitals engaged in benchmarking are advised to select 'best-in-class' organizations, little help is offered on how to do so. Even in published reports of successful benchmarking programmes, benchmark hospitals are simply described as those with 'the best rates' of performance (Barnes *et al*. 1994; Covey *et al*. 1994; Clare *et al*. 1995; Schifman 1995). Amid these various approaches and qualitative definitions, 'benchmarking is still a much misunderstood concept' (Czarnecki 1996).

In practical terms, benchmarking is used frequently to operationalize the transition from measurement of performance to performance improvement (Tomas 1993), and many quality improvement projects now present providers with benchmarks and averages to aid in this transition (Epstein 1995). The proper use of such feedback stems from a primary tenet of contemporary thinking in the field of quality – that ongoing measurement of processes and outcomes will result in continuous cycles of improvement via systematic evaluation of performance and subsequent interventions (Donabedian 1980, 1992, 1996). Such a link between measurement of quality and changes in process relies implicitly on the assumption that knowledge of one's own performance, together with the ability to compare this performance against some reference level (internal or external), will motivate improvement.

This entire process begins with the identification of well-accepted, evidence-based guidelines. Quality indicators are derived from these guidelines, and performance is described relative to these indicators. Benchmarking, then, provides targets for improvement and promotes analysis and emulation of those achieving 'best-in-class' status. Ultimately, this improvement will be reflected in better performance as measured by process and outcome indicators.

The achievable benchmark of care (ABC™)

Along these lines we have been refining our concept of the ABC™, which provides a target that is based on data derived from assessing actual performance rather than on subjective opinions as to what 'should' be done. The ABC™ encourages providers to strive for superior performance knowing that the target level of excellence has already been achieved by a select group of their colleagues. We take care to avoid a 'cookbook' approach, recognizing that

© 1999 Blackwell Science, *Journal of Evaluation in Clinical Practice*, **5**, 3, 269–281

performance standards should be realistic (Dans 1994). Since process-of-care indicators are neither flawless nor uncontroversial (e.g. long-term beta-blockers for diabetics after acute myocardial infarction) 'perfect' benchmarks are unrealistic and probably inappropriate. Therefore, an important goal of the ABC™ system is to aid in the spread of 'best practices' by a few superior providers until they become 'average care' by the majority.

We based the development of the ABC™ upon criteria we deemed essential for sound benchmarks: (1) benchmarks must represent a measurable level of excellence; (2) benchmarks should be demonstrably attainable; (3) providers with high performance should be selected from all providers in a predefined manner based on an objective measurement of actual performance; (4) all providers with high performance should contribute to the benchmark; (5) providers with high performance but small numbers of cases should not unduly influence the benchmark levels; and (6) as continuous improvement takes place, benchmark performance levels should increase.

We anticipate that several factors will generate ongoing interest in the ABC™. These reasons include the following: (1) the desire of providers to improve; (2) continuing external pressures from insurers, employers and managed care companies for increased effectiveness at reduced cost; (3) growing public demand for health care sector accountability; (4) the needs of individual hospitals and clinics to compare performance *among* organizations in order to differentiate themselves from their competitors; (5) the need to implement improvements to satisfy accreditation bodies; and (6) the need to identify superior performers so that their practices can be understood and emulated.

ABC™ methodology

Our ABC™ method can be used with any quantifiable quality-of-care indicator. However, as we explain below, we have designed the method for process of care indicators and we urge that it be applied to outcome measures with only extreme caution, if at all.

The derivation of an ABC™ begins with the total number of eligible patients for whom a specific service is 'appropriate' (denominator), and the number of those patients who received the specific service (numerator). Ideally, the criteria for 'appropriate' include the following: (1) that the process is indicated for a particular patient: for example, thrombolysis for acute transmural myocardial infarction; (2) that the process is not contraindicated: for example, beta-blockers for severe asthmatics; (3) that there is no physical reason why the intervention cannot be given: for example, a foot examination for double amputees; and (4) that the patient does not refuse treatment.

Central to the ABC™ method is what we call the 'pared mean', defined as the average performance of the subset of those providers with the highest scores for the indicator under consideration. The subset includes the top-ranked providers down to the point whereby at least 10% of the patient pool across all providers is selected. Thus, the first step in calculating the pared mean consists of ranking all providers in descending order. In previous work we explored various methods of ranking and concluded that our current approach, described below, performs best (Kiefe *et al.* 1994). Because samples with providers having a small number of cases can skew the results, this process is not trivial and demands a more sophisticated approach than simply ranking providers using crude performance.

Small denominators

Before moving to the actual benchmark calculation, we illustrate this complexity with a discussion of the small denominator problem. Consider, for example, whether physician X, who recommended mammography for her only eligible patient (100% performance), should be ranked higher than physician Y, who recommended mammography for nine of her 10 eligible patients (90% performance). The measured performance of the former physician is perfect, but the performance of the latter physician reveals a *consistently* high, but not perfect, performance. Furthermore, physician Z, ordering mammography for 90 of 100 eligible patients, might provide better insight into how she achieved excellent care than the physician with 10 eligible patients, although both performed at 90%. Therefore, for the purposes of benchmark determination, perhaps the ranking should be Z, Y, X.

A large proportion of providers with 100% performance for a small number of patients could artificially inflate the benchmark level. Data pooling, that is, summing the patient numerators and dividing by the sum of the patient denominators, somewhat attenuates this problem. In contrast, simply averaging utilization rates for the top-performing providers, instead of pooling, assigns undue weight to providers with few cases. Pooling effectively maintains providers with small denominators and excellent performance in the analysis, yet assigns lower weight to their contribution to the benchmark. More complex statistical methods of combining different top hospitals' performances are, of course, possible (Christiansen & Morris 1997). However, the quality process-of-care indicator concept presupposes enough homogeneity among providers to strive for a common benchmark. Given this assumption, we see no advantage in using a way of combining different providers' performances which would be less transparent than simple pooling.

The ABC™ methodology further addresses the small denominator problem with a correction based on Bayesian estimators which reduces usefully, but does not eliminate, the impact of providers with small numbers of eligible patients (Agresti 1990; Kiefe *et al.* 1998). This correction produces the Adjusted Performance Fraction (APF), which is calculated as follows:

$$\mathrm{APF} = (x + 1)/(d + 2),$$

where x is the actual number of patients receiving the intervention, and d is the total number of patients for whom the intervention is clinically appropriate.

The APF is lower than the actual performance for values above 50% and higher for values below 50%. For example, in the case of a provider with only one appropriate patient for whom the intervention was actually given (100% performance), the Adjusted Performance Fraction is $[(1 + 1)/(1 + 2)] = 0.67$, a difference of 33% from the actual performance. As the number of appropriate patients (d) increases, the performance fraction calculated using the ABC™ method and the percentage calculated without this correction tend to the same number, for example a provider treating eight out of 10 patients will have a Bayesian estimator-adjusted performance of 0.75 vs. an unadjusted percentage performance of 0.80.

In addition to the advantages of reducing the effect of performance percentages based on small numbers, the Bayesian adjustment allows *all* data to be used, rather than simply eliminating providers with small numbers. Finally, *the APF is used solely for the purpose of ranking providers for ABC™ calculation.* The Bayesian numbers are *not* used in the further calculations of the ABC™. In no event should these numbers be used to compare performance of individual providers.

ABC™

We now describe the calculation of the ABC™ in more detail (Table 1). First, we calculate the APF for each provider and rank order all providers in descending APF order. To establish the ABC™ denominator, we begin with the number of eligible patients for the highest ranked provider and, in descending order, add, in block, the eligible patients for subsequent providers so as to establish a provider subset that contains at least 10% of all patients across

Table 1 Steps in calculating benchmark with paired-mean method*

1. Based on the Bayesian adjusted performance fraction (APR), rank order providers (e.g., hospitals, physicians, other levels of aggregation) in descending order of performance for a specific quality indicator
2. Begin with the best-performing provider and sequentially add providers in descending order until a subset of providers that represents at least 10% of all patients or subjects in the entire dataset is established
3. Retain APF ranking, but base Step 4 on crude data
4. Calculate benchmark based on subset as follows:

$$\frac{\text{(Total number of patients in subset receiving recommended intervention)}}{\text{(Total number of patients in subset)}}$$

* Adapted from Table in Allison *et al.* 1999

© 1999 Blackwell Science, *Journal of Evaluation in Clinical Practice*, **5**, 3, 269–281

all providers. Finally, based upon the actual performance data (not the APF) calculate the ABC™ from the subset derived above according to the following formula:

$$\text{ABC}^{\text{TM}} = \frac{\text{(Total number of patients receiving recommended intervention)}}{\text{(Total number of patients in subset)}}$$

When we reach a point at which 10% of total patients are included, we call that the benchmark break-point. However, the pared mean includes all patients from all providers whose performance is tied at the breakpoint. Therefore, even though we set out to have a break-point at 10% of patients we could, after including all tied providers, have a break-point, for instance, of 12% or even higher. We recommend that the break-point be expressed to at least two significant digits. Our target is to include at least 10% of patients, but in practice it is usually a little higher. Also, note that the performance of some providers whose patients *are in the top 10%*, and therefore contribute to the benchmark, can actually fall *below* the benchmark.

The 10% target level is somewhat arbitrary. Why 10%? Perhaps because our academic histories lead us to associate 90% or above with 'A'-grade performance. The logic may be no stronger than that. Applied science abounds with arbitrarily drawn, but generally accepted, end-points: for example, the traditional 5% level of acceptable alpha error and the 80% acceptable level of beta error. Depending upon the circumstances, these conventions for alpha and beta error are sometimes modified; similarly, the ABC™ method is flexible and other users could, if they wish, select a different cut-off percentage. However, the specific cut-off should be agreed upon before benchmark performers are identified.

Aggregate measures

Whether, and how, multiple indicators could be combined into a composite indicator is a complex and relatively unexplored issue. For example, consider a composite quality indicator for a population of patients with acute myocardial infarction (AMI) where *several* different processes of care are recommended, for example administration of thrombolytic therapy, use of beta blockers, aspirin and angiotensin-converting enzyme inhibitors, and risk factor assessment. Each intervention is appropriate for a different sub-population of patients, and this heterogeneity may prevent meaningful aggregation. Also, aggregation raises the issue of weighting. For each of these interventions we have good data describing the associated mortality advantage that may be used in weighting (Mant & Hicks 1995). For other common situations where quality is measured, however, such data do not exist. Nonetheless, the public is exposed to high-profile global assessments of quality, as in the search for the nation's best hospitals (Boscarino & Chang 1997; Green *et al.* 1997; Longo *et al.* 1997; Anonymous 1998a).

Applications of the ABC™

Proper application of the ABC™ demands familiarity with our intended purpose in developing the concept, knowledge of what makes a high-quality indicator, consideration of the implications of measuring process vs. outcome and an understanding of the strengths and limitations of various data sources. Detailed discussion of these latter concepts falls beyond the scope of this paper but, because of their importance to benchmarking, we present an abbreviated discussion.

Intended use of the ABC™

Current techniques of quality improvement mandate a non-punitive approach with improvements being developed at the point of care rather than being imposed by a non-clinical authority (Shapiro *et al.* 1993). A benchmark performance level on a process indicator is a tool to improve performance, rather than a tool for the analysis of whether or not the provider is a statistical outlier. Only if providers were ranked with subsequent tangible consequences (e.g. penalties for low rank), would addressing potential random variation be important. In that situation, application of rigorous statistical methodology often leads to the conclusion that meaningful ranking is difficult, if not impossible (Bagust 1996; Andersson *et al.* 1998; Nutley & Smith 1998).

Furthermore, the medical community harbours considerable anxiety about the incorrect application of quality reports. Blumenthal & Epstein describe the generalized physician discomfort that appears to be surfacing in regard to such profiling with lack of methodological rigour (Blumenthal & Epstein 1996). It is no coincidence that the dramatic increase in profiling has 'coincided with the commercialization of the medical market place' (Berwick 1996).

As such, benchmarking may be used to improve cost of care by eliminating 'excessive' services rather than to improve quality of care (Schroeder 1995). Wennberg, noting that medical necessity does not explain the high geographic variability seen in the surgical rates in the United States, questioned: 'Which rate is right?' (Wennberg 1986; Wennburg & Cooper 1998). For procedures backed with ample evidence regarding their appropriate use, variation tends to be less. In situations where the proper indications are less clear, such as surgery for back pain, the wide-area variation is larger (Weinstein *et al*. 1998). Exploration of the higher use of bypass surgery in Ontario compared to New York led the investigators to conclude that there is 'probably neither one "right" rate nor a simple relation between service rates and appropriateness of case selection' (Tu *et al*. 1997). Therefore, we oppose the use of the ABC™ in situations where the right rate is not known, particularly when the purpose is to decrease use of medical services.

Frequently, profiling of medical care brings clinicians, who desire to do the best for their patients, into conflict with administrators, who must focus on financial accountability. Schneider describes in *Healthcare Informatics* such a situation in a large hospital in Philadelphia. Here, because of the 'culture clash' between clinicians and external consultants, participation of the medical staff in quality improvement programmes based upon practice profiles has been minimal (Schneider 1998).

Therefore, although measuring physician performance for compensation purposes abounds (Hanchak & Schlackman 1995), we discourage the application of the ABC™ to this end. Neither can we concur with ABC™s being used for credentialing purposes or for public report cards, where quantification of statistical confidence and the role of chance assume greater importance.

Indicators

The goal of clinical process-of-care indicators consists of capturing accurately the essence of clinical guidelines based on the best available evidence in a quantitative fashion, allowing large amounts of clinical data to be processed for the purpose of improving delivery of medical care (Harr *et al*. 1996; Turpin *et al*. 1996; Hofer *et al*. 1997). The following criteria, presented in priority order, guide indicator selection: (1) strength of supportive evidence (evidence obtained from randomized controlled trials will be given greatest emphasis); (2) potential to improve clinical outcomes; (3) ability to capture recommendation reliably and reproducibly during data collection; and (4) the expected frequency of application. Quality indicators suitable for benchmarking include: those measuring the use of preventive services (such as those in Healthy People 2010), those measuring appropriate management (including tests and interventions), and those measuring time to event as part of prevention or management (such as the percentage of patients with acute myocardial infarction who have an EKG within 20 min of presentation).

Process vs. outcome

We describe below our rationale for designing the ABC™ method for process of care rather than outcomes indicators (Brook *et al*. 1996).

Sensitivity vs. specificity While process measures are more sensitive to changes or disparities in quality of care, outcome measures are more specific. For example, it may take many episodes of poor operative technique to produce one adverse outcome. As a corollary, the lack of bad outcomes does not necessarily guarantee quality of care because an insufficient number of observations may circumvent the detection of an adverse outcome that would ultimately occur as more patients were treated with poor process.

Mant & Hicks performed a sample size calculation to determine the relative amount of data needed to detect a difference in quality of care between two hospitals for patients with acute myocardial infarction (Mant & Hicks 1995). Risk-adjusted mortality at 30 days served as the outcome measure, and use of aspirin, beta-blockers and thrombolytics served as the process measures. Extensive data from randomized clinical trials link each of these process indicators with a significant and quantifiable effect upon mortality in the treatment of acute myocardial infarction. Given the assumptions made in this paper, the detection of a 10% relative reduction in mortality requires 3619 patient observations. However, measuring quality of care as adherence to the process indicators requires only 48 observations to detect a 10% difference in mortality based upon the extrapolation procedure described above.

Risk adjustment Outcome measures require risk adjustment, and all methods developed to date have been less than perfect (Epstein 1995; Blumenthal & Epstein 1996). Multiple factors extraneous to the control of a particular provider impinge on outcomes, and these factors may not be amenable to risk adjustment. Wide-spread outcry led HCFA to abandon the public release of risk-adjusted hospital mortality data (Associated Press 1993). Numerous articles and reports argue that because of the inadequacy of risk-adjustment, the mortality data provided were an inadequate measure of hospital quality (Dubois 1990; Anonymous 1991; 1993; Ash 1996).

In a recent article, Palmer explains why risk-adjustment assumes less importance for process measures (Palmer 1997). Because 'process measures match patients to specific health care processes that are indicated for given conditions' they, in effect, contain 'built-in' risk adjustment. Here, additional risk adjustment may obscure important findings. For example, consider the process-based indicator that measures compliance with influenza vaccination recommendations for diabetic patients. Adjusting for age will obscure any differences in use according to age. However, influenza vaccination is more important in the elderly than in the young.

A survey of cardiovascular surgeons in Pennsylvania revealed that many were less willing to operate on severely ill patients after the institution of a risk-adjusted mortality reporting system (Malcolm 1992; Blumenthal & Epstein 1996). A similar programme in New York led to 'gaming the system' (Epstein 1995). After implementation there was a dramatic change in coding practices, making the patients appear more ill with more comorbidities. Even with these limitations data still demonstrate that mortality dramatically decreased after the introduction of such programmes.

Framed differently, risk adjustment provides a unique perspective on the epidemiological concept of confounding. Outcome measures provide indirect measures of quality of care because so many factors, in addition to quality of care, influence resultant outcomes. This mandates the application of risk adjustment which is usually performed using the epidemiological tools that address confounding. In contrast, process measures provide more direct measures of quality. When quality of care is measured by the proportion of a population receiving care that is appropriate and indicated, confounding is much less important because important exclusionary factors are reflected in the denominator of eligible patients.

Untoward manipulation Process-of-care measures may be gamed, leading some to question whether they produce better 'quality of measurement or quality of medicine'. In this spirit, the provider might design an intervention for increasing adherence to a process without influencing outcomes. Epstein notes that report cards may lead to 'focusing on the components of care that are assessed in the report card and ignoring others' (Epstein 1995). Other literature suggests specific approaches to improving process measures without necessarily improving quality of care (Clarke *et al*. 1993; Stout 1994; Parisi & Sulfaro 1996).

Data

Making an informed choice regarding data source requires, at a minimum, an understanding of the strengths and limitations of competing options. Chart audits, while being relatively expensive, provide rich clinical detail and can be tailored to the specific project at hand. Administrative data, existing mainly for billing and tracking purposes, can often be adapted to a quality improvement project (Iezzoni 1997). Use of administrative data decreases costs but raises concerns about accuracy and clinical relevance.

Examples

National benchmarks for mammography

The derivation of benchmark rates for mammography provides a concrete example where national data yield a goal that differs from those that are consensus-driven. Application of the pared-mean method to data from the National Health Interview Survey (NHIS) gave a benchmark of 71% for mammography screening, compared to the subjectively determined 60% goal of Healthy People 2000 (Allison *et al*. 1998).

Similarly, in Alabama we found from claims data that, on average, 36% of our women received mammograms and the pared mean benchmark was 72% (Kell *et al*. 1999). This benchmark means that the best-performing Alabama physicians in our study were obtaining mammography on 72% of their elderly Medicare patients. The group of best-performing physicians was defined objectively and specifically for this quality indicator using data for all study physicians. The results reflect the practices of these doctors as they struggle with the day-to-day delivery of care. Somehow, they have learned to work within the system to achieve this performance. The tenets of quality improvement hold that this accomplishment can be understood and exported. Should we aim for less?

Cooperative Cardiovascular Project

The Cooperative Cardiovascular Project (CCP), which seeks to improve quality of care for Medicare beneficiaries with acute myocardial infarction is the centrepiece of the Health Care Quality Improvement Program (HCQIP) (Jencks & Wilensky 1992). The pilot CCP, carried out in Alabama, Connecticut, Iowa and Wisconsin, became the template for an ongoing national effort (Ellerbeck *et al*. 1995). As part of the CCP, hospitals receive feedback based upon validated process of care indicators derived from the guidelines of the American Heart Association and the American College of Cardiology for the management of patients with acute myocardial infarction (Ryan *et al*. 1996). These indicators provided the basis for anonymous feedback from which hospitals were to design their own quality improvement projects with minimal outside 'interference' (Hayes & Ballard 1995).

We based our initial benchmarking work upon the pilot CCP data (Kiefe *et al*. 1994). For the Alabama pilot study, chart abstraction was performed for 3299 admissions over 107 hospitals from June 1992 to February 1993. All cases included satisfied clinical confirmation criteria for acute myocardial infarction. Focusing on five indicators (aspirin during hospitalization, aspirin at discharge, low-dose heparin, beta-blockers, ACE-inhibitors and smoking cessation counselling) we explored various approaches to benchmarking. The most robust technique, which became the foundation for the ABC™, produced high and realistic goals for all of the indicators, as shown in Table 2. Note that simply taking the 90th percentile of performance produced goals of 100% for two indicators. Also, none of the 107 Alabama hospitals performed above average on all six process-of-care indicators and, consequently, no hospital was a benchmark performer on all algorithms.

The Alabama Quality Assurance Foundation

Table 2 Data from the Alabama Pilot Cooperative Cardiovascular Project*

Algorithm	*Eligible patients*	*Hospitals caring for eligible patients*	*Average rate (95% CI)*	*90th percentile*	*ABC™*	*Hospitals above ABC™*
A	1253	106	0.14 (0.12–0.16)	0.39	0.49	10
B	502	89	0.34 (0.30–0.39)	0.83	0.73	10
C	1354	97	0.65 (0.62–0.67)	1.00	0.96	15
D	2415	105	0.57 (0.57–0.59)	0.82	0.84	8
E	370	67	0.27 (0.22–0.32)	0.75	0.67	7
F	300	50	0.51 (0.45–0.57)	1.00	0.91	9

* Adapted from Table in Kiefe *et al*. 1999.

© 1999 Blackwell Science, *Journal of Evaluation in Clinical Practice*, **5**, 3, 269–281

(Alabama PRO) successfully incorporated benchmarking methodology into the Pilot CCP Project's feedback to hospitals with positive hospital responses to that feedback (Wallace *et al.* 1996). Allowing for limitations in the non-randomized methodology, the pilot CCP documented significant improvement in several key indicators (Marciniak *et al.* 1998). Based upon the success of the pilot, HCFA expanded the CCP to a national programme. Other states now use the ABC™ as part of CCP feedback (Anonymous 1998b). We have further applied our ABC™s to several quality improvement projects, and found them to be useful, as well as generally well accepted.

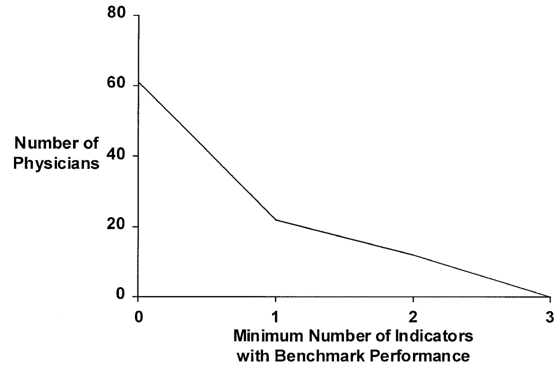

Figure 1 Cumulative benchmark trends. From the Alabama Ambulatory Care Quality Improvement Project 1997 baseline data.

Ambulatory Care Quality Improvement Project

The Ambulatory Care Quality Improvement Project (ACQIP), another component of the HCQIP, focuses on the outpatient management of diabetes. So far, ≈ 6000 medical records from physician offices in three states (Alabama, Iowa and Maryland) were abstracted and performance on five quality indicators has been analysed and fed back to individual physicians. These indicators are presented in Table 3.

Figure 1 shows baseline Alabama physicians' performance across multiple indicators. No physicians were at or above benchmark for three or more indicators. We did not provide feedback for one indicator measuring performance of dilated fundoscopic examinations, where the benchmark was ≈ 30%. Low benchmarks flag the following important possibilities that are not mutually exclusive: (1) the indicator is flawed or not accepted as a standard in the local community; (2) the algorithm for operationalizing the indicator is insensitive, missing cases where the intervention was delivered due to, for example, poor choice of variables or improper coding; (3) the data on which the indicator is based are inaccurate or do not capture the requisite information; or (4) there is a large system deficiency creating low performance. For this indicator we found that patients were receiving eye examinations from ophthalmologists, but the exams were not recorded by the primary physician.

Figure 2 shows a typical physician feedback report. Note the peer-based comparison (state average), the separate feedback for each indicator and the ABC™. All reports were individualized and each physician received only her personal data. Also, note the benchmark of 100% for creatinine measurement. We generally hope to avoid perfect benchmarks to allow for the complexity of clinical practice and so we now use the APF described earlier. However, in this situation the benchmark may be appropriate,

Table 3 Selected indicators for the Ambulatory Care Quality Improvement Project*

Receipt of influenza vaccine
Receipt of foot examination in office
Measurement of serum creatinine
Measurement of lipid profile
Determination of long-term glucose control (haemoglobin A1c, glycosylated haemoglobin)

* Numerator is number of patients receiving intervention at least once over past 12 months; denominator is all eligible patients with confirmed diabetes mellitus.

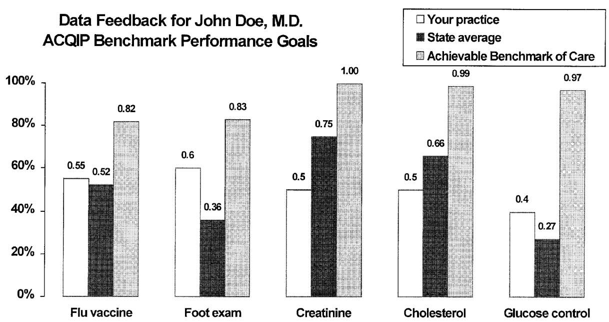

Figure 2 Ambulatory Care Quality Improvement Project feedback.

because the test is so easily ordered and obtained in most physicians' offices. If the population average is close to a perfect benchmark there is obviously little room for improvement, and resources might be better devoted to other indicators.

Provider satisfaction

A physician survey revealed high satisfaction with the above ACQIP performance profiles using the ABC™ method. We sent questionnaires to the 47 physicians participating in the Alabama Ambulatory Quality Improvement Project who received ABC™-based feedback. All physicians were surveyed ≈ 10 weeks after feedback, and the response rate was 81%. In general, the physicians rated the feedback format highly (4.42 on a five-point scale).

The questionnaire addressed several individual aspects of the report, such as the perceived accuracy of the data, the clarity of the graphs and the appropriateness of the indicators. Of all these components, physicians rated the ABC™ portion highest (4.30 on a five-point scale). Although national benchmark data is readily available from several sources (Anonymous 1995), the survey also indicated that while most physicians were interested in quality improvement generally, they are most interested in their own data and how it compares to their peers.

The ultimate test

Any quality improvement project or technique must eventually be measured against the following yardstick: did the intervention improve both process of care and patient-relevant outcomes in the most cost-effective manner (Nash 1992, 1995)? For reasons discussed previously, it is usually not practical to conduct randomized trials to determine the effect of quality improvement projects on outcomes. Instead, we rely on solid clinical evidence linking process with outcome, and assume that improving in the 'everyday' setting will improve outcomes in such a way as it did in the more strictly controlled randomized trials (Palmer 1997).

Improving process of care means changing provider behaviour. Studies do show that providing feedback of provider-specific data, either retrospectively or continuously, does, at least in the short term, change behaviour modestly (Rhyne & Gehlbach 1979; McDonald *et al*. 1984; Winickoff *et al*. 1984; Meyer *et al*. 1991; Mayefsky & Foye 1993; Cohen *et al*. 1994; Nelson *et al*. 1995; Oxman *et al*. 1995; Goebel 1997). In our own work, as described above, we use the provider profile approach for feedback and dissemination through confidential letters to the providers. We have found that acceptance of this type of feedback depends, in large part, on the perceived validity of the measures (Wallace *et al*. 1996). Presenting provider profiles within an objective *measure of performance and achievable* benchmark framework confers significant face as well as content validity to the profiles.

We are currently in the midst of several experiments to determine whether feedback of provider profiles with ABC™s improves process of care more than feedback of profiles without ABC™s. We expect a positive ABC™ effect but, if these studies are negative, the ABC™ method will not be invalidated. The ABC™ was developed to bring methodological rigour to one particular tool in the quality improvement armamentarium. As such, the ABC™ method must be judged in the same context as the entire modern quality-improvement movement as it seeks to transform evidence into practice (Berwick 1989).

Conclusion

We present the Achievable Benchmark of Care as a practical tool for a wide variety of quality improvement projects. The ABC™ possesses many advantages over the traditional approach to benchmarking. In particular, it conveys high face validity, transparency and simplicity of use. The ABC™ is derived in a predetermined, objective and standardized manner. We believe that the ABC™ best motivates positive change when used in provider feedback that is private and non-threatening, based on multiple patients, presents individual and aggregate data and offers peer-based comparisons. The derivation of global quality scores and benchmarks is an important topic for future research.

Acknowledgements

This work was supported by grant no. HS09446 from the Agency for Health Care Policy and Research. Based on a presentation at the Agency for Health Care Policy and Research meeting 'Translating Evidence into Practice '98', Washington, DC, July 1998.

References

Agresti A. (1990) *Categorical Data Analysis*. John Wiley & Sons, New York, NY.

Allison J.J., Kiefe C.I. & Weissman N.W. (1999) Can data-driven benchmarks be used to set the goals of Healthy People 2010? *American Journal of Public Health* **89**, 61–65.

Andersson J., Carling K. & Mattson S. (1998) Random ranking of hospitals is unsound. *Chance* **11**, 34–39.

Anonymous (1991) Explaining variations in hospital death rates. *Journal of the American Medical Association* **265**, 458–460.

Anonymous (1993) Dicing with death rates. *Lancet* **341**, 1183–1184.

Anonymous (1995) 1998 benchmarking guide. Data with density. *Hospitals and Health Networks* **72**, 21–30.

Anonymous (1998a) *CCP Benchmarking Provides Achievable Goals for Hospitals.* Texas Medical Foundation, Austin, TX.

Anonymous (1998b) HCIA, Mercer name top 100 hospitals, update benchmark practice patterns. *Health Care Strategic Management* **16**, 4–5.

Ash A. (1996) Identifying poor-quality hospitals with mortality rates. Often there's more noise than signal. *Medical Care* **34**, 735–736.

Associated Press (1993) Ratings of hospitals is delayed in an effort for stronger data. *New York Times* **A:** 19.

Bagust A. (1996) League tables. *British Journal of Hospital Medicine* **55**, 369–370.

Barnes R., Lawton L. & Briggs D. (1994) Clinical benchmarking improves clinical paths: experience with coronary artery bypass grafting. *Joint Commission Journal on Quality Improvement* **20**, 267–276.

Berkey T. (1994) Benchmarking in health care: turning challenges into success. *Joint Commission Journal on Quality Improvement* **20**, 277–284.

Berwick D.M. (1989) Continuous improvement as an ideal in health care. *New England Journal of Medicine* **320**, 53–56.

Berwick D.M. (1996) Quality of health care. Part 5: Payment by capitation and the quality of care. *New England Journal of Medicine* **335**, 1227–1231.

Blumenthal D. & Epstein A.M. (1996) Quality of health care. Part 6: The role of physicians in the future of quality management. *New England Journal of Medicine* **335**, 1328–1331.

Boscarino J.A. & Chang J. (1997) Commentary: inaccurate data on the quality of care may do more harm than good – an alternative approach is required. *American Journal of Medical Quality* **12**, 196–200.

Brook R.H., McGlynn E.A. & Cleary P.D. (1996) Quality of health care. Part 2: Measuring quality of care. *New England Journal of Medicine* **335**, 966–970.

Camp R. (1994) *Benchmarking: finding and implementing best practices*. ASQC Press, Milwaukee, WI.

Camp R.C. & Tweet A.G. (1994) Benchmarking applied to health care. *Joint Commission Journal on Quality Improvement* **20**, 229–238.

Christiansen C.L. & Morris C.N. (1997) Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* **127**, 764–768.

Clancy C.M. (1997) Ensuring health care quality: an AHCPR perspective. Agency for Health Care Policy and Research. *Clinical Therapeutics* **19**, 1564–1571.

Clare M., Sargent D., Moxley R. & Forthman T. (1995) Reducing health care delivery costs using clinical paths: a case study on improving hospital profitability. *Journal of Health Care Finance* **21**, 48–58.

Clarke A., McKee M., Appleby J. & Sheldon T. (1993) Efficient purchasing. *British Medical Journal* **307**, 1436–1437.

Cohen S.J., Halvorson H.W. & Gosselink C.A. (1994) Changing physician behavior to improve disease prevention. *Preventive Medicine* **23**, 284–291.

Covey B., DeVecchio J. & Mitchell L. (1994) Combining benchmarking with other quality tools to achieve breakthroughs in admitting. *Joint Commission Journal on Quality Improvement* **20**, 555–561.

Czarnecki M.T. (1996) Benchmarking: a data-oriented look at improving health care performance. *Journal of Nursing Care Quality* **10**, 1–6.

Dans P.E. (1994) Credibility, cookbook medicine, and common sense: guidelines and the college. *Annals of Internal Medicine* **120**, 966–968.

Donabedian A. (1980) *The Definition of Quality and Approaches to its Assessment. Explorations in quality assessment and monitoring.* Health Administration Press, Ann Arbor, MI.

Donabedian A. (1992) The role of outcomes in quality assessment and assurance. *Quarterly Review of Biology* **18**, 356–360.

Donabedian A. (1996) Evaluating the quality of care. *Milbank Quarterly* **44**, 166–203.

Dubois R.W. (1990) Inherent limitations of hospital death rates to assess quality. *International Journal of Technology Assessment in Health Care* **6**, 220–228.

Ellerbeck E.F., Jencks S.F., *et al.* (1995) Quality of care for Medicare patients with acute myocardial infarction. A four-state pilot study from the Cooperative Cardiovascular Project. *Journal of the American Medical Association* **273**, 1509–1514.

Epstein A. (1995) Performance reports on quality – prototypes, problems, and prospects. *New England Journal of Medicine* **333**, 57–61.

Goebel L.J. (1997) A peer review feedback method of promoting compliance with preventive care guidelines in a resident ambulatory care clinic. *Joint Commission Journal on Quality Improvement* **23**, 196–202.

Green J., Wintfeld N., Krasner M. & Wells C. (1997) In search of America's best hospitals. The promise and reality of quality assessment. *Journal of the American Medical Association* **277**, 1152–1155.

Hanchak N.A. & Schlackman N. (1995) The measurement of physician performance. *Quality Management in Health Care* **4**, 1–12.

Harr D.S., Balas E.A. & Mitchell J. (1996) Developing quality indicators as educational tools to measure the implementation of clinical practice guidelines. *American Journal of Medical Quality* **11**, 179–185.

Hayes R.P. & Ballard D.J. (1995) Review: feedback about practice patterns for measurable improvements in quality of care – a challenge for PROs under the Health Care Quality Improvement Program. *Clinical Performance and Quality Health Care* **3**, 15–22.

Hofer T.P., Bernstein S.J., Hayward R.A. & DeMonner S. (1997) Validating quality indicators for hospital care. *Joint Commission Journal on Quality Improvement* **23**, 455–467.

Iezzoni L.I. (1997) Assessing quality using administrative data. *Annals of Internal Medicine* **127**, 666–674.

Jencks S.F. & Wilensky G.R. (1992) The health care quality improvement initiative. A new approach to quality assurance in Medicare. *Journal of the American Medical Association* **268**, 900–903.

Kell S.H., Allison J.J., Brown K.C., Weissman N.W., Farmer R. & Kiefe C. (1999) Measurement of mammography rates for quality improvement. *Quality Management in Health Care* **7**, 11–19.

Kiefe C.I., Weissman N.W., Allison J.J., Farmer R.M., Weaver M. & Williams O.D. (1998) Identifying achievable benchmarks of care (ABCs): concepts and methodology. *International Journal for Quality in Health Care* **10**, 443–447.

Kiefe C.I., Woolley T.W., Allison J.J., Box J.B. & Craig A.S. (1994) Determining benchmarks: a data driven search for the best achievable performance. *Clinical Performance and Quality Health Care* **2**, 190–194.

Longo D.R., Land G., Schramm W., Fraas J., Hoskins B. & Howell V. (1997) Consumer reports in health care. Do they make a difference in patient care? *Journal of the American Medical Association* **278**, 1579–1584.

Malcolm J.A. Jr (1992) Plans proceed to publish physician-specific data. *Pennsylvania Medicine* **95**, 18–19.

Mant J. & Hicks N. (1995) Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *British Medical Journal* **311**, 793–796.

Marciniak T.A., Ellerbeck E.F., *et al.* (1998) Improving the quality of care for Medicare patients with acute myocardial infarction. *Journal of the American Medical Association* **279**, 1351–1357.

Mayefsky J.H. & Foye H.R. (1993) Use of a chart audit: teaching well child care to paediatric house officers. *Medical Education* **27**, 170–174.

McDonald C.J., Hui S.L., *et al.* (1984) Reminders to physicians from an introspective computer medical record. A two-year randomized trial. *Annals of Internal Medicine* **100**, 130–138.

Meyer T.J., Van Kooten D., Marsh S. & Prochazka A.V. (1991) Reduction of polypharmacy by feedback to clinicians. *Journal of General Internal Medicine* **6**, 133–136.

Mohr J.J., Mahoney C.C., Nelson E.C., Batalden P.B. & Plume S.K. (1996) Improving health care, Part 3: Clinical benchmarking for best patient care. *Joint Commission Journal on Quality Improvement* **22**, 599–616.

Nash D.B. (1992) Is the quality cart before the horse? *Journal of the American Medical Association* **268**, 917–918.

Nash D.B. (1995) Quality of measurement or quality of medicine? *Journal of the American Medical Association* **273**, 1537–1538.

Nelson E.C., Batalden P.B., Plume S.K., Mihevc N.T. & Swartz W.G. (1995) Report cards or instrument panels: who needs what? *Joint Commission Journal on Quality Improvement* **21**, 155–166.

Nutley S. & Smith P.C. (1998) League tables for performance improvement. *Journal of Health Services Research and Policy* **3**, 50–57.

Oxman A.D., Thomson M.A., Davis D.A. & Haynes R.B. (1995) No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. *Canadian Medical Association Journal* **153**, 1423–1431.

Palmer R.H. (1997) Process-based measures of quality: the need for detailed clinical data in large health care databases. *Annals of Internal Medicine* **127**, 733–738.

Parisi L.L. & Sulfaro M. (1996) Six steps to maximizing HEDIS results. *Journal of Nursing Care Quality* **11**, 8.

 © 1999 Blackwell Science,

Rhyne R.L. & Gehlbach S.H. (1979) Effects of an educational feedback strategy on physician utilization of thyroid function panels. *Journal of Family Practice* **8**, 1003–1007.

Ryan T.J., Anderson J.L., *et al.* (1996) ACC/AHA guidelines for the management of patients with acute myocardial infarction: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Acute Myocardial Infarction). *Circulation* **94**, 2341–2350.

Schifman R. (1995) Strategies for quality management in clinical microbiology. *Clinical Laboratory Medicine* **15**, 437–446.

Schneider P. (1998) How do you measure success? *Healthcare Informatics* March, 45–56.

Schroeder S.A. (1995) Cost containment in U.S. health care. *Academic Medicine* **70**, 861–866.

Shapiro D.W., Lasker R.D., Bindman A.B. & Lee P.R. (1993) Containing costs while improving quality of care: the role of profiling and practice guidelines. *Annual Review of Public Health* **14**, 219–241.

Stout N.J. (1994) Understanding HEDIS (Health Employer Data Information Set) can help providers attract health plan contracts. *Health Care Strategic Management* **12**, 14.

Tomas S. (1993) Benchmarking: a technique for improvement. *Hospital Material Management Quarterly* **14**, 78–82.

Tu J.V., Naylor C.D., Kumar D., DeBuono B.A., McNeil B.J. & Hannan E.L. (1997) Coronary artery bypass graft surgery in Ontario and New York State: which rate is right? Steering Committee of the Cardiac Care Network of Ontario. *Annals of Internal Medicine* **126**, 13–19.

Turpin R.S., Darcy L.A., *et al.* (1996) A model to assess the usefulness of performance indicators. *International Journal for Quality in Health Care* **8**, 321–329.

Wallace R.G., Farmer R.M., Craig A.S. & Kiefe C. (1996) Hospital responses to a data-driven statewide quality improvement project. *Clinical Performance and Quality Health Care* **4**, 34–37.

Weinstein J.N., Goodman D. & Wennberg J.E. (1998) The orthopaedic workforce: which rate is right? *Journal of Bone and Joint Surgery – American Volume* **80**, 327–330.

Wennberg J. (1986) Which rate is right? *New England Journal of Medicine* **314**, 310–311.

Wennburg J. & Cooper M. (1998) *The Dartmouth Atlas of Health Care*. American Hospital Publishing, Chicago, IL.

Winickoff R.N., Coltin K.L., Morgan M.M., Buxbaum R.C. & Barnett G.O. (1984) Improving physician performance through peer comparison feedback. *Medical Care* **22**, 527–534.