The outcomes utility index: will outcomes data tell us what we want to know?

ELIZABETH A. McGLYNN

Center for Research on Quality in Health Care, RAND, Santa Monica CA, USA

Abstract

Purpose. To propose an approach to the evaluation of the utility of an outcome measure for use in making comparisons among health plans, hospitals, networks, or other accountable entities.

Method. Seven components are recommended for evaluation: whether the outcome is a health outcome; the extent to which expectations for performance can be defined; the role medical care plays in achieving the outcome; the relative complexity of events that produce the outcome; the degree to which attribution can reasonably be made; the suitability of risk adjustment for limiting external sources of variation; and the likelihood that the measure provides perverse behavioral incentives.

Results. Illustrative examples are given in each area for scoring the performance of measures on the component.

Conclusion. Outcomes measurement is of great interest to a variety of potential users. The approach proposed here is intended to provoke discussion and more rigorous development of tools that will help to identify the measures that are likely to produce the most useful information for making comparisons among accountable entities in the health system.

Keywords: accountability, health system performance, outcomes measurement

As accountability for the delivery of high quality health care services is embraced by public and private purchasers and consumers, the standards by which those services will be evaluated have become an increasingly popular topic of debate. There are those who believe that process measures are the best way of obtaining information that stakeholders in the system need to make choices and to improve the delivery of care. There are others who believe that outcome measures are the only sensible approach because they provide the ‘bottom line’ answer to what we are buying. A more reasonable strategy is to select those measures that meet the needs of the particular application best: sometimes these will be process measures and sometimes these will be outcomes measures.

The purpose of this article is to propose an approach to the evaluation of the utility of outcomes measures for use in making comparisons among health plans, networks and providers. The thesis of this article is that the utility of any measure occurs on a continuum, from not useful to well suited to the problem at hand, and that by quantifying the utility of a variety of potential measures, we can select those that seem most likely to contribute valuable information. It is important to note that the utility of a measure for the purpose discussed here may not predict its utility for other applications. For example, the rate of low birthweight births is commonly used to evaluate population health but, because adequate risk adjustment models do not exist for this measure, it is less useful for making comparisons among health plans. Measures that are used to produce results that allow for entities to be compared fairly must meet a relatively high standard. When measures fail to meet these standards then the information produced is likely to send more noise than signal into the marketplace. At best such information will create confusion; at worst it will result in bad decisions being made that affect the health of the population negatively. The approach presented here represents the first step in developing a tool to evaluate outcomes measures for use in making comparisons among accountable entities. Additional work is required to fully operationalize this approach and to establish the reliability and validity of scoring algorithms. This article is intentionally restricted to the use of outcome measures to assess performance of the health delivery system. Although similar outcome measures may be used to evaluate the effect of other programs (e.g.
education, social services), such applications are beyond the scope of this article.

The components of the index

The Outcomes Utility Index (OUI) provides a method for assessing whether an outcome measure is likely to be useful for making comparisons among entities. The OUI has seven components each of which has a different weight in the index reflecting the relative importance of the various factors. Many of the items have scoring thresholds; measures scoring below a certain point on the rating scale are probably not worth pursuing in their existing form. The seven components (and the maximum points available) are:

- whether the outcome is a health outcome (5);
- the extent to which expectations for performance can be defined (5);
- the role medical care plays in achieving the outcome (50);
- the relative complexity of events that produce the outcome (5);
- the degree to which attribution can reasonably be made (20);
- the suitability of risk adjustment for limiting external sources of variation (10);
- the likelihood that the measure provides perverse behavioral incentives (5).

Each of these will be discussed in more detail along with the method for scoring performance on the component.

Is this a real health outcome?

As outcomes have become the favored method for assessing health system performance we have witnessed a phenomenon that might be called ‘outcome creep’ (similar to diagnosis related groups creep which referred to the practice of coding the reasons for hospital discharge in a way that maximized revenues). Donabedian, who proposed the conceptual framework which still dominates quality assessment, defined outcomes as ‘a change in patients’ current and future health status that can be attributed to antecedent health care’ [1]. He went on to explicitly include mortality, morbidity, functioning, and patient satisfaction as outcomes. The definition provides several important ideas that will be discussed throughout the index, but for this component the main point is that outcome measures must assess a change in health status.

In contrast, structural measures assess whether the characteristics of the health delivery system (e.g. availability of physicians, insurance status, use of gatekeeper arrangements) are likely to facilitate or inhibit the provision of high quality care; process measures evaluate the technical and interpersonal competence of interactions between patients and providers. There should be no penalty attached to choosing a measure of structure or process if those measures are most likely to produce information that can be used for the intended purpose.

The scoring for this component ranges from 0 to 5 with 0 representing measures that are clearly not outcomes measures and 5 representing measures that are direct (not proxy) measures of health status. Anchoring the scale may be clearer if some examples are provided. A measure such as the 30-day mortality rate following coronary artery bypass (or any other) surgery anchors the top end of the scale along with measures such as low birthweight births, or change in physical or emotional functioning. At the other end of the scale are structure and process measures such as the specialist to population ratio (structure), any measure of utilization (process) and expenditures (effectiveness). In the middle are measures that mix health outcomes with other concepts. For example, the average time to return to work following back surgery might be rated a 3 to reflect that workers compensation, the importance of work in the patient’s life, and the physical demands of the job might affect this decision as much or more than the skill of the surgeon, the ability to manage pain, and the quality of the rehabilitation regimen. Similarly, satisfaction with care is an outcome according to Donabedian but it is one that is determined by other factors such as personality, culture, and geography; as such it might be rated a 2.

The modifier ‘intermediate’ should be used with caution to reflect changes in biological status that affect subsequent health outcomes rather than to avoid using the word ‘proxy’. So, for example, this system would classify blood pressure or change in average blood pressure a 4 on this component of the outcome index. Alternatively, immunizations would be considered a process measure and would score a 0 on the outcomes scale even though the link between this process and the subsequent outcome (prevention of infectious disease) is solidly established.

Measures failing to score at least a 2 on this component should not be considered further on this index. Those are most likely to be structure or process measures that should be evaluated according to criteria that determine the utility of those measures.

Can good versus bad performance be quantified?

The second component of the OUI considers whether one can define the expectations for good performance. Expectations represent a benchmark that can be used to provide direction (such as in the Healthy People 2000 goals for the nation [2]) and can help to determine distance from the goal. The most familiar example of this comes from presentations of adjusted mortality rate data. Most surgeries have some risk associated with them and so we do not expect the mortality rate following surgery to be zero. However, the calculations of whether a procedure is worth doing should take into account the relative risks of mortality versus the benefits of the
What is the role of medical care in producing this outcome?

This component reflects that for quality assessment we are interested in outcomes that ‘can be attributed to antecedent health care’. This concept is also frequently referred to as the link between process and outcomes. The question that must be asked of each outcome measure is ‘how much of the variation in outcome can be explained by differences in the quality of medical care?’ There are several ways in which we can gain insights into this question. The first is to look at the natural course of disorder – what happens to people who receive no medical care for this problem? Alternatively, one can ask what steps the medical care system can take to improve outcomes. Are there effective methods for preventing the disease from occurring? Are there effective methods to diagnose disease early enough that outcomes are likely to be better than if the disease were diagnosed later? Are there effective methods for managing the disease so that functioning can be maintained for a longer time period at a higher level? These questions are frequently addressed in randomized controlled trials which provide the desired evidence base for many aspects of medicine. Other research designs may also contribute to drawing inferences about the likely impact of medical care.

One also should examine what magnitude of difference is possible with exceptional care, average care and below average care. If studies have large enough sample sizes, statistically significant differences in the effect of medical care interventions might be found, but those differences might not be clinically meaningful. Finally, as a way of balancing the role of medical care with other factors, one should consider what other factors influence outcomes. For example, what role do demographic factors such as age, race and sex play? Do attitude and personality of the patient contribute significantly to observed outcomes? Are there economic incentives that influence observed behavior? Do patient preferences differentially affect expected outcomes?

How complex is the process–outcome relationship?

The previous factor explored the relationship between processes and outcomes of care. This component takes the analysis one step further by asking how large and how varied is the bundle of services that constitutes the medical care procedure. The expected risk of mortality at an institution reflects the case mix of patients receiving care from that entity. Being able to quantify what we would expect to see then helps us classify what we observe as worse than expected, about right, or better than expected.

This component is also scored on a scale ranging from 0 (we don’t know what to expect) to 5 (we know and can provide norms or goals). At the high end are measures such as mortality rates following coronary artery bypass graft surgery, for which good risk adjustment models exist [3], and stage of breast cancer at detection, for which a considerable body of epidemiological evidence exists to establish population norms [4]. Many would also argue that we know enough about changes in functional status at the population level that establishing benchmarks is possible today. At the far end of the scale there is a variety of measures of disease burden or risk (e.g., the proportion of persons with asthma or diabetes, the proportion of persons who smoke) where we do not have much science to predict what we would expect from a good health plan, hospital or doctor. In the middle are measures such as complication rates following surgery (because many of them have not been well studied) and satisfaction with waiting time for a routine health maintenance visit (because individual expectations vary widely and there is very little evidence that waiting for a check-up affects health outcomes within a very large band of time).
input to the outcome being measured. Medical care can be
a simple single intervention (a ‘flu shot), a complex single
intervention (bypass surgery), an intervention with short
duration (antibiotics for a bacterial infection), an intervention
with a finite duration (antidepressants for an episode of
depression), or multiple complex interventions over a long
time period (management of insulin dependent diabetes). In
general, the longer the time that elapses between the medical
care intervention and the observed outcome, the more difficult
it is to determine with certainty what caused the outcome.
Time offers the opportunity for intervening factors to explain
as much, if not more, of the variation in outcomes as
explained by medical care.

This component is scored as the average of two individual
scores each with 5 possible points. Complexity is rated from
0 to 5 with 5 representing simple medical care interventions
and 0 representing extremely complex interventions. Duration
is rated from 0 to 5 based on the time between the initiation
of medical interventions for the problem and when the
outcome is observed, with 5 representing little delay between
process and outcome and 0 representing time frames longer
than 5 years. For example, low birthweight would rate a 0
for complexity and a 3 for duration for an average score of
1.5. Post-surgical infection would rate a 4 for complexity and
a 5 for duration for an average score of 4.5. Amputation of
a limb as a complication of diabetes would rate a 0 for both
complexity and duration. Blood pressure control would rate
a 4 for duration and a 3 for complexity for a score of 3.5.
Antibiotics for an upper respiratory infection would rate a 4
for both complexity and duration for an average score of 4.

**Is attribution of responsibility for outcome reasonable?**

When outcome measures are used as the basis for comparing
the performance of two or more entities, the implication is
that the entity bears responsibility for achieving the outcome.
This requires that a plausible link exists between the entity
and the means by which the outcome is achieved, that the
entity has had an adequate amount of time to influence the
outcome, and that, where appropriate, shared responsibility
for producing outcomes is acknowledged (e.g. taking the role
of personal choice into account).

This component has a score of 20 possible points ranging
from 0 (‘I have no control over anything in this measure’) to
20 (‘If I were the measured entity I would be delighted
to be evaluated on this outcome’). Table 1 illustrates some
possible scores on this component for different health system
entities.

Mortality rate following bypass surgery is an example of a
measure that has been sufficiently well developed to be used
to hold hospitals accountable. The primary care physician
might be held partially responsible for making an appropriate
referral; the medical group to which the surgeon belongs
has a higher level of responsibility and the managed care
organization (because it makes hospital contracting decisions)
is just below the hospital. By contrast, the low birthweight
rate is not well understood in terms of the entity responsible
for producing better outcomes. At the hospital level, it is too
late in the provision of care to modify the birthweight
(although survival rates for such infants might be monitored),
responsibility at the level of individual doctors or groups is
difficult to assign, and even at the managed care organization,
in the absence of an adequate risk adjustment model, as-
ignment of responsibility is difficult. Glycemic control is an
example with fairly high rates of attribution to three of the
providers; some element of patient responsibility for
adherence to recommended therapy is reflected by assigning
responsibility scores lower than 20. Some responsibility for
mortality following auto accidents is assigned to physicians,
groups and managed care organizations because active pro-
grams to counsel about use of seat belts may affect the
likelihood of survival. Hospital measures may be more reason-
able, although regional comparisons may be more valid than
national comparisons because rates are likely to be affected
by the adequacy of the trauma system in the area.

**Is risk adjustment adequate?**

In most cases, multiple factors contribute to outcomes. In
quality measurement, we are interested in isolating the
component that relates to the medical care system, which is
accomplished by controlling for the other factors that con-
tribute to the outcome. Risk adjustment strategies must,
however, be parsimonious rather than exhaustive. Factors
that are reasonable to exclude from risk adjustment models
are those that occur rarely, explain relatively little of the
variation in outcomes, or are randomly distributed among

---

### Table 1 Sample results for evaluation of outcome measures on attribution of responsibility (maximum points = 20)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Primary care MD</th>
<th>Physician group</th>
<th>Hospital</th>
<th>Managed care organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality rate post CABG</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Low birthweight</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Patient satisfaction</td>
<td>15</td>
<td>10</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>Glycemic control</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Mortality from auto accident</td>
<td>5</td>
<td>2</td>
<td>15</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 2  Assessment of the adequacy of risk adjustment for comparisons among health plans (maximum points = 10)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality following CABG</td>
<td>8</td>
</tr>
<tr>
<td>Glycemic control</td>
<td>7</td>
</tr>
<tr>
<td>Improvement in symptoms of depression</td>
<td>5</td>
</tr>
<tr>
<td>Low birthweight rate</td>
<td>2</td>
</tr>
<tr>
<td>Substance abuse relapse rate</td>
<td>1</td>
</tr>
</tbody>
</table>

The outcomes utility index

the evaluated entities. Factors that are frequently included in risk adjustment models include patient characteristics (age, sex), severity of the illness that is the focus for measurement, and comorbid conditions.

Ten points are allocated to this component and scoring ranges from 0 (no risk adjustment model and external factors swamp medical factors in determining the outcome) to 10 (all of the important external factors have been taken into account). Table 2 shows some illustrative measures with scores on this component.

The risk adjustment model for mortality following coronary artery bypass graft is adequate at the level of hospital and surgeon and by extension would be reasonable at the health plan level. Similarly, glycemic control reflects good management of diabetes and health plans that provide adequate counseling and other services to support physicians should be able to achieve high performance on such a measure. Some may argue that patient compliance should be taken into account, but good systems and providers should be able to affect compliance. Effective treatments exist for depression but the risk adjustment models are not as well developed. The substance abuse relapse rate is complex and no adequate risk adjustment models are available; the disorder itself is characterized by relapse and research has not established clear process–outcomes links or identified risk factors for likelihood of success or failure in a single episode of treatment.

Conclusions

Considerable interest in using outcome measures to evaluate health system performance has been expressed by a variety of stakeholders in the system. Using outcome measures is valuable if the information conveyed is valid for making comparisons among evaluated entities. Unfortunately, there are a variety of challenges in using outcome measures appropriately for this purpose. Unless a measure addresses the potential pitfalls adequately, the results produced from the measure may be misleading rather than helpful.

The OUI provides an approach to evaluating potential outcomes measures that is intended to stimulate discussion and development of appropriate evaluation tools by the outcomes research field. Since outcomes measures are likely to range from extremely useful to worse than no information, the index (once fully developed) could be used to identify the most promising measures for use by organizations responsible for making external comparisons. Quality measurement should not be viewed as simply an academic exercise. Measurement and reporting should stimulate the health care system to provide better services which in turn result in improved health. To realize this goal of quality measurement we must have tools to select objectively those measures that get us closer to our goals as quickly as possible.

References

1. Donabedian A. The Definition of Quality and Approaches to its Assessment. Ann Arbor, MI: Health Administration Press, 1980.


Accepted for publication 31 July 1998