Using health outcomes data to compare plans, networks and providers

R. HEATHER PALMER

Center for Quality of Care Research and Education, Harvard School of Public Health, Boston, MA, USA

Abstract

Purpose. To analyze the challenge of using health outcomes data to compare plans, networks and providers.

Analysis. Different questions require different designs for collecting and interpreting health outcomes data. When evaluating effectiveness of treatments, tests or other technologies, the question is what processes improve health outcomes? For this purpose, the strongest evidence comes from a double-blind randomized controlled trial. In program evaluations, the question is ‘what is the impact of this policy and related programs on health outcomes?’ For this purpose, we may be able to randomize subjects, but are more likely to have a quasi-experimental or an epidemiological design. When we compare plans, networks and providers for quality improvement purposes the question is ‘do these specific plans perform differently from one another?’, or, ‘are these specific plans improving their performance over time?’ We want to isolate for study the effects attributable to specific plans. Designs that yield strong evidence cannot be applied because we lack experimental control.

Conclusions. When we already have strong evidence linking specific processes of care with specific outcomes, comparing process data may reveal more about performance of plans, networks and providers than comparing outcomes data. Comparisons of process data are easier to interpret and more sensitive to small differences than comparisons of outcomes data. Outcomes data are most useful for tracking care given by high volume providers over long periods of time, targeting areas for quality improvement and for detecting problems in implementation of processes of care.

Keywords: quality comparisons, quality improvement, study design

Interpreting health outcomes data

In his classic paradigm of structure, process and outcome [1], Donabedian uses the phrase ‘health outcomes’ to refer to measures of the health of patients, specifically observations of events within a persons body and mind. Health outcomes do not include events such as admissions or readmissions: these events are processes of care that can be used as proxy measures of worse health. However, finding persons who were readmitted does not necessarily find all whose health deteriorated. Some people who became ill enough to justify admission may have received such poor care that their illness went un-noticed or untreated.

By definition, health outcomes data relate to a prior intervention. We use the outcomes data to infer if an intervention caused the state of health that we observe in a group of persons.

Figure 1 represents outcomes data for three groups of persons: A, B and C. Our discussion is not limited to data samples of nine persons, however: nine circles are simply convenient for the diagram. The persons represented in this diagram are not selected by any clinical condition. Any type of outcome measure is shown here: we can imagine many possibilities, from clinical test results to patient-reported health status measures.

On observing the persons in Figure 1 after some prior intervention, we can only say that they are now different in their health. They may also have been different before the intervention, implying that the intervention had no effect. Or, if these differences in health relate to a prior intervention, it may not be the intervention that we have in mind, but some other intervention that we may not have considered.

Address correspondence to Dr R. Heather Palmer, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. Tel: +1 617 432 0779. Fax: +1 617 432 3199. E-mail: hpalmer@hsph.harvard.edu

© 1998 International Society for Quality in Health Care and Oxford University Press
Interpreting outcomes data. This diagram represents outcomes data for three groups of persons: A, B and C. Each circle represents a person. The darkness of the circle represents the degree of health of each person using an outcome measure at a point in time. A lighter circle means better health; darker means worse health. Nine circles are shown in each group as a convenience but this does not signify that the sample size was restricted to nine persons per group.

What do these data tell us? For group A, the time 2 people are less sick, and for group B the time 2 people are more sick. For group C, there is no difference in how sick the people are at time 1 and time 2.

If these data came from three plans, what could we deduce? Perhaps a group of well people joined plan A between time 1 and time 2. Perhaps a group of sick people joined plan B between time 1 and time 2. We cannot tell which explanation is correct.

If the enrollment populations were stable in all three plans, the data could reflect simply the natural history of health and disease in people who happen to belong to three different plans. Maybe plan A had many enrollees in time 1 with short-lived illnesses who were destined to get better. Perhaps some very sick people in plan A died between time 1 and time 2 and so were not sampled in time 2. Maybe plan B had many people in time 1 who had illnesses that we would expect to get worse by time 2. Perhaps some significant environmental event hit enrollees in plan B just before time 2, such as a flu epidemic, and this epidemic didn’t affect enrollees in plans A and C.

We could address these questions better by stratifying the data by diagnosis or clinical condition. This design would account at least for one source of variability in the natural history of their health. We can also distinguish causes and effects better if we collect these time series data prospectively for the same persons at two points in time—a cohort study—as illustrated in Figure 3. In Figure 3 we can see that two persons labeled A1 and A2 stayed sick whereas A3 got much better. Person B3 got much worse.

However, with this design we lose sight of persons who died between time 1 and time 2 (perhaps because they got bad care) or who left the plan between time 1 and time 2 because they believed that they got bad care.

If we select the persons in Figure 3 by diagnosis we can improve our ability to distinguish effects caused by the time...
course of different illnesses further. For instance, if all these persons had a heart attack at time 1, we might choose to set time 1 and time 2 at 2 years apart in order to observe the effects of the care given. However, for persons with high blood pressure, who are treated in order to prevent complications such as stroke, heart attack and kidney disease that can occur many years later, we must set time 1 and time 2 many years apart. We will undervalue the care given if we observe hypertension outcomes at 2 years.

Health promotion and disease prevention are also under-valued by short term outcome measurements – ‘I will not feel any better or appear any better today on any outcome measure if my provider gives me a pneumonia vaccination today, or provides a mammogram today, or persuades me to start eating less fat today, but these interventions may save my life some time from now’.

Selecting or stratifying groups by diagnosis requires confirmation that diagnoses were correctly made. Suppose we selected patients with a diagnosis of heart attack for outcomes comparison. What if providers in plan A often over-diagnose heart attacks? We might include in plan A, as heart attack victims, patients A3, A4 and A8 who only had heart burn – of course they will be less likely to have angina or heart failure by time 2. On the other hand, if plan A providers under-diagnose heart attacks, they may send patients with heart attack home from the emergency room. These unfortunate people may die without our knowing it because they would not be surveyed at time 2.

Another way to disentangle the effects of the provider from all other effects in time series comparisons is to adjust the time 2 data for the risk of a bad outcome associated with the health status of individuals at time 1. If we could ‘adjust’ for diagnosis, severity of illness, co-morbidity and other risk factors that affect the outcome, we level the odds of showing improved outcomes for all three plans. We can discern more easily any differences between the groups in their outcomes, and so make fairer comparisons.

With comparisons adjusted for risk factors, we can better determine which plan’s patients got better or worse, relative to the health status they could expect at time 2. For instance, we might find, after adjustment or stratification, that the difference in outcomes among these plans shown in Figure 3 disappears. However, risk adjustment would not help to explain why all plans got the same results. It could be that all plans give effective care in equivalent amounts. Alternatively, it could be that there is no treatment for this particular disease so that no provider could have altered the course of health for these patients. It could also be that the effective treatment is complicated to follow and so patients in all plans tend to drop out of treatment.

Similarly, if plan A got better results than plans B and C, we still would not know why. Suppose there is a new and effective treatment. Maybe the providers in plan A kept up to date and so offered this treatment whereas the providers in plans B and C were not aware that an effective treatment existed. Maybe the effective treatment is very expensive and plans B and C refuse to pay for it. By refusing to pay for drugs, plans B and C might be able to offer lower premiums so that persons with low incomes might be attracted to them. However, these persons might not be able to buy the effective treatment with their own money. We need to distinguish between these explanations of differences in health outcomes among the three plans in order to decide how to improve care.

If we look only at outcomes data, we cannot tell whether an effective treatment exists, whether it was offered to patients, and if offered, whether patients took the treatment. How can we separate these causes and effects? To answer this question we must consider methods to determine the effectiveness of treatments.

To determine whether a treatment accounts for differences in outcomes for individuals with a particular disease, we need a radical revision of the study design: Figure 4 illustrates such a design. Instead of collecting health outcomes data on people who belong to different plans, we collect data on people who did or did not receive treatment, irrespective of their enrollment in any plan.

Many patient outcomes research teams use the quasi-experimental design illustrated in Figure 4. Notice that the patients in treatment groups A and B, those persons whose providers decided to offer them treatment, are mostly sicker than those in group C who were not offered treatment. We can only begin to interpret these data inasmuch as we can control for patient variables, other than treatment, that affect outcome.

Researchers who classify study designs grade this type of study as producing weak evidence. For instance, the US Preventive Services Taskforce used in its deliberations a five-level hierarchy for strength of evidence. Expert opinion gets the lowest value for strength of evidence. Uncontrolled time series produce slightly stronger evidence. Still greater credibility attaches to well-conducted case-control and cohort
Figure 5 Comparing treatments: randomized controlled trials. Data collected for persons who were randomly assigned to groups, with groups then randomly assigned to treatment protocols, treatment protocol A, treatment protocol B or no treatment. Outcomes measurements were made using standardized instruments at standardized intervals before and after treatment/no treatment with both providers and patients blinded to treatment status. (See Figure 1 for explanation of symbols.)

The strongest evidence of cause and effect comes from believe that you have strong evidence when you see tables of health outcomes data that were not collected under the randomized controlled trials [2]. Figure 5 illustrates a randomized controlled trial (RCT).

The strongest evidence to evaluate a treatment comes from collecting outcomes data in a double blind RCT [3]. This design overcomes the problems of interpretation that we examined above. In Figure 5, groups A, B and C represent persons who have the same disease and are similar in severity and co-morbidities. We assign these persons to each group using random number tables. This usually creates groups with exactly the same mix of patients, even for those patient differences that we overlooked or cannot measure. We assign the three groups by random number tables to treatments that are given according to strict protocol. We blind both providers and patients as to the group that each patient is in. We measure health outcomes using standard instruments at standard times relative to treatment. We include adequate numbers of individuals to allow for the effects of chance.

With this design, if differences are statistically significant or exceed the confidence intervals, we believe them strongly. From the data in Figure 5, for instance, we would conclude that treatment A works well, and that treatment B works less well.

This design yields credible results, but of course we cannot apply it in comparisons of plans. When we compare plans for quality improvement purposes we cannot randomize persons to plans. With this and every other departure from the double blind RCT format, the strength of our causal inference linking providers to health outcomes falls.

People concerned with health care are used to seeing data from RCTs and believing the results. Physicians, particularly, often see tables of health outcomes data comparing results at time 1 and time 2. It is easy to forget all the special conditions that constrained the collection of these data if they came from an RCT. If you forget that, you can mistakenly believe that you have strong evidence when you see tables of health outcomes data that were not collected under these strict experimental conditions. And, of course, we usually do not have strict experimental conditions when we are comparing plans as opposed to treatments. Figure 6 illustrates how easy it is to slide into this kind of mistake.

The data format in Figure 6 looks just like the format in Figure 5; however, in Figure 6, persons in groups A, B and C are not randomly assigned to their group, and groups are not randomly assigned to treatments. Instead, Figure 6 shows persons who selected themselves to enroll in different plans and who differ in severity of illness.

Even if we selected these individuals by diagnosis, we could not confirm the diagnosis because that is decided by the physician who saw the patient. These individuals are not receiving treatment according to a strict protocol, but as they and their physician decide. Any outcome measurements made may occur at erratic intervals in non-standardized notations made by a variety of health care practitioners. Still more troubling, the health care practitioners making outcome measurements are not blinded to treatment. Even if standard outcomes measurement instruments are used for plan comparisons, such as patient health status surveys, and even if these are administered at standard time intervals, the patients themselves are not blinded to treatment. The numbers of individuals within a plan, particularly if we stratify by diagnosis, may not make up an adequate sample size to exclude chance effects.

This design doesn’t yield strong evidence; in fact, without careful interpretation the data can actually be misleading. For instance, what if plan B is giving much better care than plans...
Comparing health outcomes data

Clarifying the questions

The discussion above leads us to this conclusion: we should clarify the question we are asking before we design our outcomes data collection.

When our purpose is evaluating effectiveness of treatments or screening tests or other technologies, we are asking what processes improve health outcomes? For this question, the strongest evidence comes from a double-blind randomized controlled trial. To overcome the rigidities of this design, we may prefer quasi-experimental or epidemiological studies. Nevertheless, throughout the study we employ methods to isolate the effect of the treatment or technology, and suppress effects attributable to plans, networks or providers.

In program evaluations, we are asking what is the impact of this policy and related programs on health outcomes? For this purpose, we may be able to randomize subjects, but are more likely to have a quasi-experimental or an epidemiological design. In every aspect of study design we try to isolate the effect of the program from effects attributable to specific plans, networks or providers.

However, when we make provider comparisons our goal in study design is the exact opposite because we are asking quite different questions. Do these specific providers perform differently from one another? Do these specific providers perform satisfactorily compared with some absolute standard? A and C in both time 1 and time 2? Perhaps, because it is giving such good care at time 1, large numbers of sick people rush to join it; by time 2, plan B will look worse – but is it because sick people joined it, or because it gave such bad care that mildly ill people deteriorated and became sick?

The difficulties of making sound interpretations of data derived from outcomes comparisons make an alternative approach attractive – namely, using process data for plan comparisons. If we already have strong evidence from prior research, showing that certain patients will improve in health if given a certain treatment, we need not repeat that research in the less controlled situation of plans giving routine care. We can simply ask how plans compare on giving the effective treatment for this type of person. Figure 7 illustrates this type of design: instead of a prospective study collecting outcomes data over time, we can do a retrospective study of the prior year using process of care data.

This design has major advantages when our purpose is stimulating quality improvement or guiding purchases. It is relatively inexpensive to collect data, gives quick results, and is more sensitive to small differences between plans than an outcome study [4]. We can also use the design to compare plans on performance of tests that are shown, by scientific evidence, to improve health outcomes, such as screening tests.

In some circumstances, process comparisons may be valuable even if we don’t have strong scientific evidence [2]. Consider an airplane that crashes because a substantial number of bolts are missing from the airplanes wing. No-one has conducted an RCT to test whether omitting bolts from airplane wings causes planes to crash, but most of us would not fly in a plane that we knew had missing wing bolts.

Using process data for comparisons

In plan comparisons, when should we prefer process to outcomes data? We can use process data whenever we have strong evidence or beliefs that particular processes affect important outcomes. Process data are especially useful for comparisons when one or more of the following apply:

- the goal is improving delivery of care;
- we need to know why specific providers achieve particular outcomes;
- short time frames are necessary;
- the processes of interest affect long-term outcomes;
- performance of individual providers is of interest;
- performance of low volume providers is of interest;
- tools to adjust or stratify for patient factors are lacking;
- providers are being compared in a competitive/coercive situation.
If we are responsible for improving care, we must eventually get data on the processes of care, because the only way we can improve patients outcomes is by improving what we do for them.

If we compare providers on outcomes to identify those with the worst outcomes, then we must study their processes of care to see how they might improve. What are they doing or not doing that makes their results worse? However, even if we find all providers have similar outcomes we will still need to study processes. What if they have similar outcomes because they are all falling short in giving effective treatments to appropriate patients? The Medicare program encountered this situation in the Cooperative Cardiovascular Disease Program [5]. The study reported data on use of processes of care shown by strong evidence to reduce deaths from heart attacks, namely thrombolitics, beta blockers and aspirin. All of the study hospitals could have achieved better outcomes than they did: they failed to use these treatments for all patients without contraindications to treatment.

Process data are useful when we need results of comparisons in a short time frame, and when the processes affect important long-term outcomes. Process data are the best way to compare provider groups or individual providers who contribute only the part of the care received by a patient. In this common circumstance, when we look at outcomes only, we cannot tell which providers to credit for a good outcome, or if all should share the credit.

If the provider groups under comparison have small numbers of cases eligible for study, outcomes studies are not possible, because detecting differences in outcomes requires large sample sizes [4].

If we lack methods to adjust for differing risks for a bad outcome, process comparisons are more interpretable than outcome comparisons, although process comparisons, too, are sharpened by adjustments or stratification on relevant patient factors.

Outcome studies are especially problematic when we are using comparisons for coercive or competitive purposes. Providers in these situations have a big stake in the action that follows from the results. They may contest the findings or start gaming to evade them. If we have used a weak study design, the contest over interpretation of data can go on virtually forever. Gaming to ensure better-looking health outcomes data can produce negative effects for patients.

Using outcomes data for comparisons

In plan comparisons, when should we prefer outcomes to process data? We can use outcomes data whenever we have tools to adjust or stratify for patient factors that affect the health outcome. Outcomes data are especially useful for comparisons when one or more of the following apply:

- we seek areas for quality improvement;
- specific processes are known to yield specific gains in outcomes;
- safe implementation of processes is of interest;
- long time frames are possible;
- performance of whole systems is of interest;
- performance of high volume providers is of interest;
- comparing providers in a cooperative situation.

Suppose we want to prioritize areas for quality improvement, and have evidence about the processes of care that improve outcomes for a clinical condition, and know how good the outcomes should be if that care were given well. In this situation, outcome comparisons help us to detect whether the outcomes achieved are less good than they should be. This may be the best way to uncover mistakes and oversights in implementing care. Finding error-prone areas of care falls into the category that John Williamson calls ‘achievable benefit not achieved’ [6]. When outcomes fall short of the level known to be possible, we have found a good place to spend limited resources for quality management because there are opportunities for big improvements.

Outcomes comparisons are particularly useful when we can wait to observe long-term outcomes, and are interested in the impact of whole systems or programs of care on the health of large populations. When we compare large volume providers, we can obtain sufficient sample sizes to detect differences in outcomes.

Because the strength of evidence in non-experimental comparisons is weak, outcome comparisons are most useful when working with providers in a voluntary and cooperative mode. Using outcome comparisons in coercive and competitive situations, where each provider has high stakes in the result, can encourage gaming that produces perverse effects. For instance, providers may avoid enrolling sicker patients. By avoiding sicker patients, providers can achieve better outcomes [7]. Better outcomes can also result if risky procedures are withheld from higher risk patients [8]. Plans can appear to improve outcomes if sicker patients disenroll: just raising the barriers to care may be enough to persuade sick persons to go elsewhere [9]. Up-coding diagnoses and documentation of patient factors that are known to be used to adjust outcomes data can also mimic real improvement in outcomes [10]. Of course, improving accuracy of coding is desirable, but exaggerating the nature of the patients illness when selecting disease coding is not.

Conclusion

Using health outcomes data to compare plans, networks and providers poses many challenges that are often overlooked. In reviewing these challenges the following points emerge:

- In interpreting health outcomes data, we must clarify the question we are asking. Different questions require different designs for collecting and interpreting the outcomes data;
- Designs that yield strong evidence, and that are typically used in evaluating treatments, cannot be applied when comparing plans, networks and providers for purposes of quality improvement;
- When we have good evidence linking specific processes
Comparing health outcomes data

of care with specific outcomes, comparing process data may reveal more about performance of plans, networks and providers than comparing outcomes data.

- Outcomes data are useful for tracking care given by high volume providers over long periods of time, and for detecting problems in implementation of processes of care.

References


Accepted for publication 31 July 1998