Outcome measures assess the effects of health care on the health status of patients (Donabedian, 1988). Although the documentation of outcome measures such as mortality has long history, modern outcome assessment was not established until the 1970s as the result of at least three factors (Epstein, 1990; O’Connor and Neumann, 2006). First and perhaps most significant was the perception of variation in clinical processes from doctor to doctor, hospital to hospital, and region to region. This variation raised questions about the quality of health care, and measuring health outcomes was seen as one way to answer these questions (Epstein, 1990; Kassirer, 1993). A second contributing factor was the influence of cost containment on quality. The growth of managed care and the emphasis on efficiency led to concerns about the negative effect this might have on health care quality. Outcomes measures were seen as one way to monitor possible deterioration in the system (Epstein, 1990). Third, heightened competition among insurers also led to the perceived need for outcome measures that provide a metric of value for different health care plans. Although price plays a key role in this competition, buyers were and still are interested in the relative value of the health care they purchase.

It is difficult to exaggerate the impact that measuring health outcomes has had on contemporary health care. In terms of recent funding and impact, consider that in 2009 Congress allocated $1.1 billion of the American Recovery and Reinvestment Act for comparative effectiveness research—the goal of which is to improve health outcomes. To give a sense of the scale of this investment, the U.S. Agency for Healthcare Research and Policy (AHRQ) received $300 million of this funding, which represented a 500% increase in the Agency’s expected 2009 budget for comparative effectiveness research. Moreover, in 2010, Congress authorized the establishment of the Patient-Centered Outcomes Research Institute (PCORI) as part of the Patient Protection and Affordable Care Act. Furthermore, the U.S. is not alone in its focus on health outcomes. In the 2008 review of the UK’s National Health Service (NHS), High Quality Care for All, Lord Darzi writes, “...we can only be sure to improve what we can actually measure.” (Department of Health, 2008). The UK has since put into place a number of reforms focused on measuring health outcomes, including the NHS’s Quality and Outcomes Framework and the Patient-Reported Outcomes Measures (PROMs) Programme.

As the measurement of health outcomes plays an increasing role in the management and evaluation of contemporary health care, it has also begun to receive attention within philosophy. To be sure, concepts of health and well-being have long had a place in philosophy, but only more recently have philosophers started to engage with the measurement practices and
epistemic consequences of contemporary outcome assessment. In beginning such an engage-
ment, one of the first things to note is that terms such as “outcome,” “outcome measure,” and
“measurement outcome” are used frequently—and sometimes ambiguously—in the medical
and philosophical literature. It is thus useful to be clear about what these different terms mean.

When authors refer to “outcomes” or “health outcomes,” they are usually referring to what
Eran Tal calls the object of interest (Tal, 2015). The primary outcomes or objects of interest
in medicine and surgery are mortality (how many people die after or during receipt of health
care), morbidity (a population’s disease load as the result of health care), and quality of life
(patients’ perceptions of their health-related quality of life). These outcomes are measured using
a variety of instruments; for instance, quality of life is typically measured using self-reported
questionnaires, and morbidity may be inferred from cholesterol assays or clinician-reported
questionnaires. In the medical literature these instruments are sometimes referred to as “out-
come measures” (e.g., quality of life is often measured using questionnaires collectively referred
to as Patient-Reported Outcomes Measures, PROMs). Outcome measures are then put to use
within a study (i.e., descriptive or analytic), the aim of which is to answer a question about
a health outcome(s) in a particular context. Within the context of a particular study design,
the results gathered via these instruments are reported using as a statistic, e.g., relative risk
(comparison of risk of, e.g., mortality, for different groups of people), difference between means
(subtraction of two group averages, e.g., the mean quality of life score of two populations), or
Quality Adjusted Life Year (QALY, a measure of disease burden over time). These results, or
what are sometimes called “measurement outcomes,” are the knowledge claims about the val-
ues of one or more quantities attributed to the outcome of interest (Tal, 2015).

In what follows I begin with a look at how measures of mortality and morbidity are used
in clinical research. To illustrate the role that outcomes, outcome measures, and measure-
ment outcomes play, as well as some of the philosophical questions that they raise, I turn to
the UK’s Place of Birth study (Birthplace in England Collaborative Group, 2011). After my
consideration of mortality and morbidity, I turn to quality of life. As will become clear, the
kinds of philosophical questions that concern quality of life measures are somewhat different
from those that frame philosophical interest in mortality and morbidity. In part this difference
is because quality of life measures are a more recent addition to outcome measurement and in
part it is because quality of life is a more complex outcome.

Mortality and Morbidity

In the U.S. and UK, births in obstetric units vastly outnumber births that take place outside
of obstetric units. Still, non-obstetric births are increasing in both countries. For example, in
2004 only 0.87% of U.S. births occurred in non-obstetric units (home or midwifery units),
but by 2012 1.36% occurred in a non-obstetric unit. In England and Wales, they have seen an
even steeper increase, with only .9% of births occurring at home between 1985–1988, rising to
2.4% in 2011 (Office for National Statistics, 2013). Is it professionally responsible to support
a non-obstetric birth? Should clinical guidelines support choice in place of birth? These kinds
of questions shape the debate over place of birth. These questions tend to be answered—and
debated—via empirical studies investigating the relative riskiness in terms of mortality and
morbidity of different birthplace choices.

Take for example the Birthplace in England Research Programme, which in 2011 published
the largest prospective cohort study of its kind, looking at 64,500 births to women at low
risk for complications, including almost 17,000 women who planned to give birth at home
(Birthplace in England Collaborative Group, 2011; Hollowell, 2014). The Birthplace in Eng-
land Research Programme was a multidisciplinary research program.
coming out of this research, “Perinatal and maternal outcomes for planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study,” compared three non-obstetric unit locations (i.e., home, free-standing midwifery units, and midwifery units alongside obstetric units) with obstetric units to determine whether outcomes in the non-obstetric units differed from those in the obstetric group. Based on the study’s results, the authors conclude that both healthy low-risk women who have never before given birth to a live baby (nulliparous women) and healthy low-risk women who have given birth to at least one live baby (multiparous women) should be offered a choice of birth setting (Birthplace in England Collaborative Group, 2011).

**Primary Outcome**

The primary outcome or object of interest in this study was a mixture of neonatal morbidity and mortality. Specifically, researchers were interested in measuring the following morbidity and mortality events: stillbirth after start of care in labor, early neonatal death (within seven days of birth), disturbed neurological functioning (neonatal encephalopathy), meconium aspiration syndrome (when a baby breathes in a mixture of amniotic fluid and feces near the time of delivery), brachial plexus injury (damaged shoulder nerves resulting in loss of movement or weakness of the arm), and fractured humerus or clavicle (Hollowell, 2014). A mixed outcome of mortality and morbidity events was used instead of focusing solely on a single mortality or morbidity event in order to give the study more statistical power.

Statistical power refers to a study’s ability to detect a difference between two groups if a difference between them really exists. In this study, power refers to the study’s ability to detect differences among planned places of birth. When the differences you are searching for are relatively rare, as they are in the case of the above adverse events, then it is difficult for a study to detect genuine differences with regard to the rare event. One way to increase a study’s power is to combine rare events and thus increase the probability of their occurrence in the study. Nonetheless, the choice to use a mixed outcome in this study has been controversial.

Lachlan de Crespigny and Julian Savulescu (2014) write that, as a result of this choice, the study is “unhelpful” in determining the relative risk involved in choosing to give birth at home. They argue that it is unhelpful because the outcome includes disparate conditions (e.g., perinatal mortality, encephalopathy, humeral and clavicle fractures), which vary in the seriousness of their prognosis. For instance, while hypoxic-ischemic encephalopathy (HIE) contributes to long-term disability, fractures are typically less debilitating. Their argument is that without evidence of the relative risk of individualized morbidity outcomes (e.g., HIE), we cannot properly assess the safety of home births.

De Crespigny and Savulescu go on to argue that because children born at home can expect a delay in diagnosis, delivery, and/or transfer following an acute intrapartum event, it is more likely that homebirths lead to a higher rate of HIE events and subsequent disability (de Crespigny and Savulescu, 2014). Jennifer Hollowell and Peter Brocklehurst, authors of the Birthplace study, respond to this claim arguing that while mixed outcomes can obscure differences in outcomes among birth sites, it is speculation to say that the risk of long-term disability is higher at home. Indeed, although the confidence intervals are wide, the Birthplace study does not suggest an excess of neonatal encephalopathy in multiparous women planning a homebirth: neonatal encephalopathy occurs in 1.2 events per 1,000 planned homebirths (99% CI .6 to 2.2) vs. 1.8 per 1,000 planned obstetric unit births (99% CI .8 to 3.7) (Hollowell, 2014). Confidence intervals (e.g., 99% CI .6–2.2) provide an estimate of uncertainty associated with the statistical estimate of a population parameter (e.g., mean, odds ratio, weighted incidence).
De Crespigny and Savulescu respond to this point arguing that the Birthplace study does not provide us with information regarding the differences in long-term avoidable disability according to place of birth. This is both because the study was not powered to individuate the primary outcomes and because it did not include a long-term follow-up. Consequently, we cannot conclude anything about the relative riskiness or safety of homebirth. Another way to put de Crespigny and Savulescu’s argument is that the logic of the Birthplace study is invalid: the conclusion does not follow from the premises.

Measuring Instrument

In the above paragraphs I discussed the outcomes that the researchers in the Birthplace Study were interested in measuring; now I move on to discuss the instrument through which they acquired information about these outcomes. In this study information regarding the primary outcome was collected via forms completed by the midwife who started intrapartum care and in some cases with the help of a member of the clinical team on the admitting neonatal unit. Epistemologically speaking, these data collection forms are complex and in part serve as the measuring instrument for the primary outcomes (in part they also serve to provide other kinds of information, e.g., demographics of the study). Consider the fifth and sixth questions in Section E of the Home Birth Data collection form (Birthplace in England Collaborative Group, 2011):

“Was this baby admitted to a neonatal unit within 48 hours of birth?” yes □ no □
“If yes, to where was the baby admitted?” □ Special Care Baby Unit □ High Dependency Unit □ Neonatal Intensive Care

Date baby was discharged from neonatal unit: □ □/□□/□□
Not yet discharged □

Were any of the following identified in the baby within 48 hours after birth?

□ Meconium aspiration syndrome □ Neonatal encephalopathy □ Fractured humerus □ Fractured clavicle □ Fractured skull □ Neonatal sepsis □ No morbidity identified □ Admission to neonatal unit within 48 hrs of birth for at least 48 hrs with evidence of feeding difficulties or respiratory distress □ Other morbidity Please specify________

Tal makes a useful distinction between measurement indications and measurement outcomes, which help mark the transition between answers to questions such as those above and the results of the study, which I discuss below (Tal, 2013). Measurement indications describe states of the measuring instruments (e.g., answers to questions), and measurement outcomes refer to knowledge claims about quantities attributed to the object being measured (e.g., morbidity).

The measurement indications in this example are the positively marked tick boxes that a midwife checks to indicate answers to the questions posed on the data collection forms. Measurement outcomes are acquired by statistical analysis of multiple indications, often taking into account missing data and systematic error. With regard to the latter in the Birthplace study, neonatal encephalopathy required clinical review and coding to ensure accurate and
consistent treatment. For example, a clinician reviewed each data form in which there was no confirmed diagnosis of neonatal encephalopathy, but there was a record of isolated seizures (which might indicate encephalopathy). If there was no cause of seizure other than presumed oxygen deprivation, then neonatal encephalopathy was coded as the outcome (Birthplace in England Collaborative Group, 2011).

**Measurement Outcome**

Measurement outcomes of mortality and morbidity are typically given as a statistic. It is not uncommon for these statistics to represent some kind of risk. Risk can be expressed as absolute or relative risk. Absolute risk refers to an individual’s chances of an adverse outcome over a period of time (e.g., intrapartum period). In the Birthplace study, absolute risk was expressed as the weighted incidence of the primary outcome per 1,000 for women planning to give birth at difference locations. Relative risk compares the risk between two groups of people. In the Birthplace study, relative risk was expressed as an odds ratio between, for example, nulliparous women planning a homebirth and nulliparous women planning an obstetric birth.

As both Jacob Stegenga and Alex Broadbent have noted, it is common practice for clinical studies to report their results in terms of relative risk at the expense of absolute risk (Broadbent, 2013; Stegenga, 2015). This practice can be misleading because relative risk does not take into account the baseline rates of the outcome in question, and as a result doctors and patients tend to overestimate the risk of the intervention in question. Stegenga argues that it is less misleading for studies to present both kinds of statistics (Stegenga, 2015). But while the Birthplace study does present both absolute and relative risk, controversy regarding which statistic to emphasize continues.

Consider the results of the Birthplace study for all of the women planning obstetric and homebirths. There were 4.4/1,000 primary events in the planned obstetric unit group and 4.2/1,000 in the planned homebirth group. Looking at absolute risk, planned homebirths appear safer for healthy women with low-risk pregnancies than planned obstetric unit births, but there are different ways of dividing up the cohort of women planning an obstetric and homebirth, which can alter the interpretation of the results.

For example, part of the way through collecting data for the study, it was discovered that almost 20% of women in the obstetric unit group, compared with ≤7% in each of the other settings, had at least one complicating condition at the start of labor (e.g., high blood pressure or abnormal fetal heart rate). This difference suggested to the researchers that the risk profile was different among low-risk women in the study. The study’s steering group decided to analyze a restricted population of women who did not experience complication at the start of labor. In this cohort we see a weighted incidence per 1,000 of 3.1 for the obstetric group and 4.0 for the homebirth group. The odds of a primary event was 59% higher for those planning a homebirth.

What should be made of the findings from this study? In their conclusions the Birthplace study authors repeatedly emphasize that, “the incidence of adverse perinatal outcomes was low in all settings” (Birthplace in England Collaborative Group, 2011). In other words, the authors emphasize the statistical findings expressed in terms of absolute risk and as a result see individual choice as a reasonable policy. Some, including the UK’s National Institute for Clinical Excellence (NICE), have agreed with these conclusions, and in April 2015 NICE decided to update its intrapartum care guidelines. These guidelines now recommend that all healthy low-risk women should be given a choice of birthplace (NICE, 2014).

But others referring to the same study have interpreted the evidence differently. For example, Frank Chervenak and colleagues argue that offering women a choice of birth setting “is irrational and cannot be supported in light of the reported adverse outcomes for birth outside
of an obstetric service” (Chervenak et al., 2013). In this article Chervenak et al. focus exclusively on outcomes reported in the Birthplace study, referencing the population of women who did not experience complications at the start of labor. Moreover, they report solely the odds ratio of 1.59 to argue that the choice of a homebirth is “irrational,” and they never mention absolute risk.

What should we take from this disagreement over whether the measurement outcomes from the Birthplace study support a reasonable or irrational choice regarding place of birth? First, while it is certainly better to include absolute and relative risk in a study, doing so does not forestall disagreement over the interpretation of the results. Second, this disagreement points to philosophical issues beyond the scope of the present chapter. One of those issues is a question regarding what population cohort best represents the general population of women with low-risk pregnancies: the restricted population or the study population? Independent of absolute or relative risk, the answer to this question could help direct our attention to one set of statistics.

Another issue concerns non-epistemic values (e.g., ethical, aesthetic, prudential values) that are at play in interpreting these measurement outcomes. For instance, while the authors of the Birthplace study consider low-risk pregnancies to be a low-risk activity, de Crespigny and Savulescu look at the same mortality and morbidity data and conclude that birth is “inherently risky” (de Crespigny and Savulescu, 2014). Inmaculada de Melo-Martin and Kristen Intemann (2012) have argued that this difference in interpretation is at least partly due to the non-epistemic values that the different authors hold. Although we can only speculate, the position that Holowell and colleagues take on the riskiness of childbirth seems to be motivated by values of respect for nature, individuality, expertise (midwifery), trust, and what we might call clinical frugality. The positions that Crespigny and Savulescu, and Chervenak and colleagues take on this same issue seem to be influenced by values of respect for the unknown, duty, science, control, and restraint (McClimans, 2015). It is in virtue of holding these different values that these researchers can read the same data and understand its significance differently.

Quality of Life

Although “hard” outcome measures such as mortality and morbidity originally led the outcome assessment movement, “soft” outcome measures of quality of life were quickly taken up as part of it. The popularity of these measures—often referred to collectively as Patient-Reported Outcome Measures (PROMs)—is due to at least two factors. First, for many contemporary health interventions, particularly elective ones such as hip and knee replacements, the point is less one of mortality and morbidity and more a matter of improvements in quality of life. Indeed, even in cases such as place of birth, if the mortality and morbidity is roughly equivalent among birthplaces, then quality of life becomes a deciding factor—indeed it is a deciding factor for many low-risk women. For instance, women may choose to give birth in a non-obstetric unit because they want to avoid some of the risk factors associated with urinary incontinence. Non-obstetric units are associated with decreased probability of having an episiotomy (a cut made in the pelvic floor muscle during delivery) (Birthplace in England Collaborative Group, 2011). Episiotomies are associated with an increased probability of urinary incontinence. The quality of life associated with urinary incontinence is measured with a PROM (e.g., Urinary Incontinence Quality of Life Scale, I-QOL).

Second, patients play an essential role in the assessment of quality of life, since the measuring instruments used (i.e., questionnaires) seek answers directly from patients about their quality of life. One reason to ask patients directly about their quality of life is that third parties such as clinicians and caregivers tend to underestimate patients’ quality of life (Sprangers and
Another reason is because patient-centered measures promise to ameliorate some of the hermeneutic marginalization that occurs when patients enter the technology and measurement-heavy setting of contemporary health care (Institute of Medicine, 2001). Once again, consider the Birthplace study. Although there is controversy over how to understand the significance of the measurement outcomes associated with neonatal mortality and morbidity, there is almost zero controversy regarding the maternal morbidity measurement outcomes. Maternal morbidity is significantly lower in non-obstetric units (Birthplace in England Collaborative Group, 2011; Chervenak et al., 2013). Although studies have not yet done so, using quality of life measures to compare place of birth might help to move women and their health from the periphery of this debate.

But for all their popularity, PROMs have received significant criticism. One line of criticism focuses on the under-theorized nature of quality of life and the problems this creates when trying to develop a valid and interpretable measuring instrument. Another line of criticism involves one form that the measurement outcome can take: the quality adjusted life year (QALY). In what follows I will consider both criticisms, but before doing so it is helpful to keep in mind one basic distinction: as measuring instruments, PROMs can be either utility or non-utility measures.

Philosophers tend to be more interested in utility measures than non-utility measures. This is because quality of life measures that are associated with utility values enable the generation of QALYs. QALYs aid in priority setting for scarce health care resources and thus pose philosophical questions regarding, for instance, justice. For instance, we might wonder how place of birth impacts one’s quality of life over time. Indeed, one way to read de Crespigney and Salvulescu’s argument is that planned homebirths lead to decreased QALYs since they lead to greater disability. To be sure, this point is speculation as there is currently an absence of QALY data for these outcomes (Schroeder et al., 2012).

Utility measures are always index measures; this means that a patient’s quality of life score (e.g., the sum of the instruments indications) can be added together to produce a single number. In the context of a utility measure, this score is correlated with utility values that are derived from the public’s ranking of different health states to yield a measurement outcome. The public ranks health states using thought experiments such as the time trade-off technique. The time trade-off technique provides individuals with two scenarios, and they are asked which scenario they prefer. The length of time in the second scenario is varied until the individual is indifferent between the two. Traditionally, the first scenario offers a medical intervention, which results in a chronic condition, $i$, that lasts for 10 years followed by death. The second scenario offers a medical intervention, which results in full health for $x < 10$ years followed by death. When the value of $x$ is such that an individual is indifferent between the two scenarios, then the utility value for $i$ is given by $x/10$. The EQ-5D (EuroQol) is an example of a utility measure of quality of life that uses the time trade-off technique (Brooks, 1996).

There are more non-utility measures than utility measures (Measuring Health, 2006). The main difference between them is that non-utility measures cannot generate QALYs because they are not associated with utility values. As a result, non-utility measures are not used in conjunction with priority setting in health care, and they do not raise the same philosophical questions about justice. Instead, non-utility measures are used in other contexts (e.g., as endpoints in trials to determine the clinical effectiveness of different medical interventions, to support the claims made on the labels of pharmaceuticals, and to help determine the quality of the health care patients receive).

Non-utility measures can be either profile or index measures (i.e., producing a single score or multiple scores for each questionnaire). A profile measure is developed when the construct under measurement (i.e., quality of life of those with cancer) is deemed to consist of multiple
independent measurement scales (e.g., physical functioning, emotional role functioning, mental health). Two examples of non-utility measuring instruments are the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ) and the aforementioned Urinary Incontinence Quality of Life Scale (I-QOL).

Problems with Quality of Life Measurement: Interpretability

The first line of criticism regarding the under-theorized nature of quality of life typically refers to non-utility measures. This criticism can be summarized by a single deficiency that has multiple consequences—namely, quality of life measurement generally lacks the theoretical resources to provide a representation of the measurement interaction (i.e., the relationship between the primary outcome or construct, e.g., quality of life and its instrument, i.e., a PROM) (Van Fraassen, 2010). This criticism is not uncommon amongst psychologists and others working within the field. Consider Donna Lamping’s (2008) Presidential address to the International Society of Quality of Life Research, in which she identified the need for a theoretical framework as one of three challenges facing the future of PROMs. Or take Jeremy Hobart et al.’s (2007) *Lancet Neurology* article, in which they lament the lack of explicit construct theories in their article criticizing the current state of PROMs (Hobart et al., 2007). Sonja Hunt, in her 1997 editorial for *Quality of Life Research*, argues that the surfeit of poorly designed measures suggest that we do not know what quality of life is (Hunt, 1997). In what follows, I provide a brief overview of one consequence that results from this lack of theory: problems with interpretability.

Over the last 15 years, the discussion of how to interpret change in patient-reported outcomes has received considerable attention. Interpretability refers to the clinical significance of increases or decreases on a particular measure over time. For instance, if I score a 30 on the Urinary Incontinence Quality of Life Scale (IQOL) directly after having an episiotomy, then we know that I have scored toward the bottom end of the scale—the IQOL has a 100-point scale, and higher scores indicate better quality of life. But imagine two months later I score 42. What does this 12-point increase mean from a clinical point of view? Should my drug regime change? If so, how should it change? Most PROMs only provide ordinal-level information; i.e., we know that someone who scores 42 is more depressed than someone who scores 30, but we do not know the degree of that difference. PROMs are thus difficult to interpret.

This difficulty has led to the development of methods to enhance their interpretability. One popular method is the identification of a minimal important difference (MID). A MID is the smallest change in respondent scores that represent clinical, as opposed to merely statistical, significance and that would *ceteris paribus* warrant a change in a patient’s care (Jaeschke et al., 1989). One method for determining a measure’s MID is to map changes in respondent answers onto some kind of control. The idea is to determine the minimal amount of change that is noticeable to patients and to use this unit of change as the MID. This method asks the control group of patients to rate the extent of their symptom change over the course of an illness or intervention on a transition-rating index (TRI). TRIs are questionnaires that ask patients questions, such as “Do you have more or less pain since your first radiotherapy treatment?” Typically, patients are given seven possible answers ranging from “no change” to “a great deal better” (Fayers and Machin, 2015). Those who indicate minimal change (i.e., those who rate themselves as just “a little better” than before the intervention) become the patient control group. The mean-change score of this group is used as the MID for the PROM.

This approach of acquiring a MID via a patient control group assumes that respondents who rate their symptom change as “a little better” on a transition question should *ceteris paribus* also have comparable change scores on the PROM. Put differently, similarities in respondent answers to transition questions ought to underwrite similarities in respondents’ magnitude of
change over the course of an intervention or illness, but qualitative data from interviews with
patients suggests that this assumption is ill-founded (Taminiau-Bloem et al., 2011, Wyrwich
and Tardino, 2006). Whether one understands the magnitude of change over the course of an
illness as large or small is a matter of interpretation.

Consider Cynthia Chauhan, a patient advocate during the deliberations on the FDA guide-
lines for the use of PROMs in labeling claims. During her testimony, she discussed the side
effects of a drug called bimatoprost, which she uses to forestall blindness from glaucoma. One
of the side effects of bimatoprost is to turn blue eyes brown. Chauhan has “sapphire blue” eyes,
in which, she says, she has taken some pride. As she speaks of her decision to take the drug
despite its consequences, she notes that doing so will affect her identity in that she will soon
no longer be the sort of person she has always enjoyed being—i.e., she will no longer have blue
eyes (Chauhan, 2007).

We can imagine that, even if the bimatoprost is only minimally successful and Chauhan’s
resulting change score from the PROM is low, she will nonetheless have experienced a signifi-
cant change—she will not be the same person she was before. But this significance is tied to
the place that her blue eyes had in her understanding of herself and what she took to be a good
life; ceteris paribus we would not expect a brown-eyed person to summarize their experience in
the same way. Thus, it would not be surprising if Chauhan’s answer to the transition question
was “quite a bit,” while the magnitude of her change score was minimal.

I suggest that what examples such as this illustrate is that our understanding of clinical sig-
nificance ought to be closely linked to our understanding of quality of life given the cohort of
respondents for whom the outcome measure is targeted (e.g., octogenarians). To put this point
slightly differently, understanding change in PROMs requires that researchers have a grip on
what quality of life means in the context of a particular PROM and the population it serves.

**Quality Adjusted Life Years**

Thus far I have been discussing quality of life in terms of difficulties with theorizing the con-
struct and the consequences this has for developing an interpretable outcome measure. The
second line of criticism regarding quality of life refers to utility measures and one statistic com-
monly used to express their measurement outcomes: the quality-adjusted life-year (QALY).

QALYs are a way of valuing and comparing health outcomes. They combine the length of
survival with a measure of the quality of that survival and assume that given a choice a person
would choose a shorter life of high quality over a longer life of poor quality. A year of life in
perfect health is associated with 1 QALY and death is associated with 0. To determine the val-
ues between a PROM, such as the EQ-5D (EuroQol), provides indications that are correlated
with utility values derived from the public’s ranking of different health states. Earlier I gave the
example of the time trade-off technique as one way to value the health states provided by the
EQ-5D (EuroQol). These respective utility values are multiplied by the number of years that
patients are expected to live in that particular state of health. This multiplication provides a
QALY between 1 and 0. QALYs can then be used to compare the quality of life of different
individuals in a population or the same individual given different treatment regimes.

Consider the following example. Since we do not have QALY data for place of birth, the
following QALY gains are fictitious but not inconceivable. Imagine a group of women giving
birth in an obstetric unit and a group giving birth at home. For the obstetric units, further
imagine that the QALY gain (before and after birth) is .0 and the QALY gain (before and after
birth) for the homebirth group is -0.01. According to these QALY scores, obstetric unit births
are associated with better quality of life than homebirths, but where should women give birth?
The answer depends at least in part (but as we saw earlier when we discussed the neonatal
mortality and morbidity measurement outcomes, only in part) on whether obstetric unit births are cost effective. The average cost of a homebirth in the UK is approximately $1,180, and the average cost of an obstetric unit birth is approximately $1,730 (Schroeder et al., 2012). To determine cost effectiveness we divide the difference of the cost of the respective locations by the difference of the QALY gains. The cost per QALY in this case is $55,000. Institutions such as the UK’s National Institute for Clinical Excellence (NICE) often set cost per QALY thresholds over which new treatments are not considered cost effective. NICE’s current threshold is at the £20,000–£30,000 mark (Appleby et al., 2007).

This example illustrates an interesting point. Even if quality of life over time is slightly worse when giving birth at home, it still may not be cost effective to encourage women to give birth in obstetric units (as we currently do in the United States). To be sure, this point begs the question of using cost per QALYs to determine healthcare priorities. One common criticism of QALYs is that they use a societal point of view instead of an individual point of view to value health states. With regard to the first point, it is not the women giving birth who determine the utility value associated with their health state. Recall from earlier that the general population determines utility values when they participate in thought experiments such as the time trade-off exercise. Thus, women giving birth provide information about their health state via a PROM such as the EQ-5D, but their score from this measuring instrument is associated with a utility value determined by the general population. This utility value is used, along with time, to generate a QALY. Is it appropriate for the general population, some of which are individuals who will never give birth, to value the different places where one can give birth?

QALYs have also been criticized on the grounds that they are ageist (Harris, 1987). Because part of the QALY algorithm requires the number of years lived in a particular health state, older members of a population are at a disadvantage simply because their natural life expectancy is less than younger members of the same population. Other criticisms of discrimination have also been leveled at QALYs. If low cost per QALY interventions translate into health care priorities (at the expense of high cost per QALY interventions), then those individuals with conditions that are relatively cheap to treat will be prioritized. This way of priority setting could introduce systematic bias if, as is in fact the case, certain people in a population require a greater investment of resources to obtain a decent quality of life.

Despite these criticisms, QALYs are the measurement outcome of choice in most resource allocation contexts. They are the cornerstone of economic analysis for governments, managed care, and other health care payers (Kind et al., 2009). Few are overly enthusiastic over QALYs, but many believe that while they are not perfect, they are useful (see, e.g., Hausman, 2006). In 2006 the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) held an invited panel entitled “Will the QALY Survive?” The panelists acknowledged various degrees of difficulties, but they agreed that it would survive. Dennis Fryback usefully summarized the reasons: (1) the QALY provides a metric for the allocation of resources that combines capacity and functioning; (2) it does a decent job indicating the average impact of an intervention on a population; and (3) we do not have an alternative.

References


31–38.
40(12), 807–812.
de Melo-Martín, I., and Intemann, K. (2012) “Interpreting evidence: why values can matter as much as
science,” *Perspectives in Biology and Medicine* 55(1), 59.
The Stationary Office.
Association* 260 1743–1748.
Epstein, A.M. (1990) “The outcomes movement—will it get us where we want to go?,” *New England
Journal of Medicine* 323 266–270.
1105.
Hollowell, J. (2014) “Homebirth and the future child: Factual inaccuracies in commentary on the Birth-
place study,” *Journal of Medical Ethics*, Letter http://jme.bmj.com.pallas2.tcl.sc.edu/content/40/12/807.
abstract/reply.
Institute of Medicine (IOM). (2001) *Crossing the Quality Chasm: A New Health System for the 21st Cen-
mal clinically important difference,” *Controlled Clinical Trials* 10 407–415.
329 1263–1265.
Oxford University Press.
O’Connor, R.J., and Neumann, V.C. (2006) “Payment by results or payment by outcome? The history of
London: Office for National Statistics
Schroeder, E., Petrrou, S., Patel, N., Hollowell, J., Puddicombe, D., Redshaw, M., and Brocklehurst, P.
(2012) “Cost effectiveness of alternative planned places of birth in woman at low risk of complic-
aions: Evidence from the Birthplace in England national prospective cohort study,” *British Medical
Journal* 344 e2292.
in evaluating the quality of life of patients with chronic disease: A review,” *Journal of Clinical Epide-
miology* 45 743–760.
medical Sciences* 54 62–71.

**Further Reading**