

STUDY DESIGNS IN BIOMEDICAL RESEARCH



DESIGN ISSUES IN
QUALITY OF LIFE STUDIES

What is HEALTH or HEALTHY?

- ▶ It might be beyond your thinking; it's more than just “not sick” or “no diseases”.
- ▶ Medicine has acquired its softer side; expanded concept of health is not new.
- ▶ The World Health Organization (WHO) defined health (in 1948) as “ a state of complete physical, mental, and social well-being and not merely the absence of infirmity and disease”.
- ▶ To reflect this focus on a broader picture of health, researchers started to put more efforts on “**Health-related Quality of Life**” - the impact of disease and its treatment on the physical, mental, and social well-being of an individual.

Health-related Quality of Life

- ▶ Traditionally, **clinical trials have focused on endpoints that are physical or laboratory measures of response**; e.g. disease-free survival.
- ▶ Those endpoints do not reflect how the patient feels and functions in daily activities, yet these perceptions reflect whether or not the patient believes s/he has benefited from the treatment.
- ▶ More recently, clinical trials are including endpoints that reflect the patient's perception of his or her well-being and satisfaction: **Health-related Quality of Life (HR-QoL)**!

Example:

A question from an instrument/questionnaire called FLIC (Functional Living Index of Cancer):

On the scale from 1 (Not at all) to 7 (A great deal), how much is pain or discomfort interfering with your daily activities?

“How a patient feels” is not physical; “How a patient functions” leads to physical measurements, but the kind of measurements not directly related to morbidity, mortality, or their treatments. These are new domains of HR-QofL research.

EARLY EFFORTS

- ▶ Simple measures of health status - such as “Karnofsky scale/score” in cancer (Karnofsky and Burchenal, 1949) - began to appear within clinical medicine in the late 1940s in response to inadequacies of mortality and morbidity as description of outcome in many diseases, especially chronic diseases.
- ▶ In the USA, in the late 1960s and early 1970s, concerns about the effectiveness as well as economic costs led the National Center for Health Services Research to support the development of measures of health status.

Chronic diseases are those persisting for a long time; patients do not die early but suffering along the way. For patients of chronic diseases, therefore, the issue is not “survival” but “to live better” – that is to have “quality time”.

“**Health Services Research**” is a field related to health policy, and health services management; its domains include the financing, organization, delivery, and outcomes of health services.


“Health-related Quality of Life” are research specifically based on “patient-reported” outcomes which refer to the extent to which one’s usual or expected physical, emotional, and social well-being are affected by a medical condition or its treatment. That, of course, raises the question of subjectivity.

Example: The first question from SF-36 (36-Item Short-Form Health Survey):

In general, would you say your health is: Excellent, Very good, Good, Fair, or Poor

OBJECTIVE VERSUS SUBJECTIVE

- ▶ There may exist a misconception that objective assessments are more valid, that “patient-reported” outcomes may not agree with those of “trained professionals” which constitute the “gold standard”.
- ▶ But many biomedical endpoints that we consider objective include measurement errors, not agreed among experts, have poor predictive values.



For example, high cholesterol values likely lead to heart diseases; but does lowering cholesterol prolong life? i.e. are what you measure the “outcome”, the “cause” or the “symptom”? It is not easy to tell if some phenomenon – something you measure – is the cause or the symptom?

And health professionals – even the best doctors - do not know how the patient feels more than the patient does about him/herself.

AREAS OF APPLICATION

- ▶ In the last two decades interest has increased in quality-of-life (Qof L) measures in four (4) broad health contexts:
 - (1) Measuring the health of populations,
 - (2) Assessing the benefit of “alternative” resources,
 - (3) Comparing two interventions in a clinical trial, and
 - (4) Making a decision on treatment for an individual patient (medical oriented, just as “diagnosis”).
- ▶ Each context requires an assessment of the impact of ill health on the aspects of the everyday life of the subject.

There are no studies conducted for HR-QoL research. HR-QoL assessments are primarily designed to compare groups of patients receiving different treatments in clinical trials, and to identify change over time within these groups of patients (context #3).

That is, in “protocol language”, “Quality of Life” issues are in “secondary aims”.

EXAM #1:

TREATMENTS FOR NEUROBLASTOMA

Advanced form of neuroblastoma, a life threatening disease in children, are aggressive cancers (with solid tumor) with generally poor diagnosis. They are often treated with a cocktail of cytotoxic drugs; severe side-effects are almost always associated with these treatment regimens. A multi-center randomized trial was conducted (Pinkerton et al., 1988) to compare a single high dose of “melphalan” versus placebo (following induction therapy).

“Quality of Life” forms a secondary outcome measure to supplement the primary outcome, overall survival time from randomization.

EXAM #2:

TREATMENT FOR MILD HYPERTENSION

Hypertension is an important risk factor for cardiovascular disease but, for most patients, it's asymptomatic (silent killer!). The benefits of the drug treatment of hypertension, primarily in reducing strokes, have been well established. However, controversy remains over risk-benefit ratio for "mild hypertension" where treatment benefits (which are delayed) may be offset by adverse effects (immediate).

The issue is which treatment produces minimal interference with a patient's "Quality of Life".

EXAM #3:

ARTHRITIS & RHEUMATISM

Rheumatoid arthritis (RA) is a chronic disease which is rarely fatal; its main health impact are: pain, stiffness, and functional limitations (hard to raise hands over head). Several laboratory, radiological, and clinical measures are used to assess the course of the disease (whether it progresses or responds to a treatment); but there is no universal agreement on which measure or combination of measures best represent the short-term or long-term outcome.

May be, “Quality of Life” measures should be explored in Rheumatology!

For patients with many cancers, the issue of “Quality of Life” is just a minor subtext in discussions of treatment choice. For patients with these “rapidly lethal” diseases, particularly those “for which cure is possible”, the goal of both physician and patient is to use “as aggressive a treatment as necessary” to achieve remission. The primary goal is patient’s survival, e.g., “5-year survival”; one worries about “life” before its “quality”.

Any toxicity during and complications after treatment are often considered “necessary evils”. But there are exceptions, especially those based on the above two key words: “rapid” and “cure”.

For “bone cancer patients”, especially those in advanced and/or terminal stage, pain becomes unbearable and QofL becomes “the issue”.

For “prostate cancer patients”, especially those in early, non-metastatic phase, the disease progresses very slowly and there may be little survival benefit of active treatment for older patients. For these patients, accurate information about the “likely complications” of treatment is required to provide the ethical and legal basis for treatment decision and for patient informed consent.

INSTRUMENTS

- ▶ In health status assessment measures, aspects of the patient's perceived well-being are self-assessed and a score is derived from the responses on a series of questions; these questions form several “domains/dimensions” about different aspects of daily life during a “recent period of time”.
- ▶ There are two basic types of instruments - generic (e.g.. SF-36) and disease specific (e.g.. Functional Living Index of Cancer FLIC, or Functional Assessment of Cancer Therapy FACT), or Breast Chemotherapy Questionnaire BCQ, etc...).

Perhaps the two most popular instruments or measures are the “Karnofsky index” (measuring the physical performance of patients with cancer on a scale of 0 to 100; which is often used as one of the inclusion and exclusion criteria) and “SF 36” a short form with 36 questions, which is not disease-specific.

The main concern is that Instruments developed in one context are sometimes applied in studies with different objectives - especially when important consequences such as treatment decisions or resource allocation depend on them.

MORE ON INSTRUMENTS

- ▶ Qof L concept started as “global & holistic”; a more reductionist approach is necessary.
- ▶ The concept is broken into components called dimensions; e.g. “Physical well-being”, “Psychological well-being, and “Social well-being”.
- ▶ Each dimension is further decomposed into a number of questions called ‘items’; which in turn may be answered on a “scale” - for example, 5-point scale, 10-point scale.
- ▶ In describing and comparing groups, or “arms” of a clinical trial, one “combines” the detailed responses back into dimensions - or into a single “global assessment”.

PSYCHOMETRIC ASSESSMENT

- ▶ For any disease group or sub-group, there are a large number of “instruments” (i.e. questionnaires); some may be suited, some may not. For example, for cancers, we have Functional Living Index of Cancer (FLIC) and Functional Assessment of Cancer Therapy (FACT)- among others.
- ▶ The starting point would be some assessment of characteristics of a specific instrument, especially their psychometric properties: Reliability and Validity.

EVALUATION OF INSTRUMENTS

The criteria for evaluation of instruments have been the concern mainly of psychologists and psychometricians. Briefly, items and scale are:

- (1) Valid,
- (2) Reliable, and
- (3) Responsive or Sensitive

VALIDITY

- ▶ “Validity” involves assessing an instrument against an “accepted absolute standard” which, in the case of QofL, is not available; One can base on weaker criteria and assessment is often more descriptive for formality.
- ▶ “Content validity” involves checking whether items appear to cover its intended topics clearly and unambiguously - by professionals in field.
- ▶ “Construct validity” involves inspection of the overall pattern of relationships between the instrument and other measures; for example, between a disease-specific instrument - such as FLIC - and SF36 for overall health.

RELIABILITY

- ▶ The “Reliability” of an instrument is a measure its “ability to yield the same results on repeated trials under the same conditions”.
- ▶ Reliability appears to be easier to assess than Validity but there are always problems because the “test-retest” approach proves to be practically difficult in the context of clinical trials: patients undergo changes in health between the test and the retest.
- ▶ The popular alternative is to examine an instrument’s “Internal Reliability” (at a single administration) using “Cronbach’s Alpha”.

CRONBACH'S ALPHA

- ▶ The “Reliability” of an instrument is a measure its “ability to yield the same results on repeated trials under the same conditions” which is not usually done.
- ▶ Two sets of measurements for the same variable may not have exactly the same value; however, they must show some “consistency” - an indication of reliability.
- ▶ Cronbach’s Coefficient Alpha” (1951) applies not to “repeated measurements” but to “interrelated items” - used as a substitute for “repeated measurements” in single administration; it measures Internal Reliability.

EXAMPLE

- ▶ The “Functional Living index-Cancer” (FLIC) consists of 5 domains and 22 items. Morrow et al. (1992) tested the entire 22-item scale and found Cronbach’s Alpha to be .89 Separately, however, the results showed that it’s highly consistent for Physical well-being and Psychological well-being (over .80) but less so for Social well-being (just over .60) Keep in mind that Cronbach’s alpha is only the “lower” bound of a Coefficient of Determination, so “.60+” is not bat at all.

EVALUATION OF INDIVIDUAL ITEMS

- ▶ To determine how each item reflects the reliability of the dimension/domain, one calculates a coefficient Alpha “after deleting that one item from the dimension,
- ▶ If the Cronbach’s alpha increases after the deletion, the deleted item is “not” correlated highly with other items; Conversely, if the Cronbach’s Alpha decreases, that item is highly related with other items in the domain.

RESPONSIVENESS

- ▶ Any QofL instrument used in a clinical trial should be able to detect clinically significant changes over time; but “Responsiveness” or “sensitivity” has been studied far less than Validity and Reliability.
- ▶ One may compare an instrument against some established “criterion of change” using ROC curve; however, there are difficulty in deciding the criterion, e.g. by doctor or patient.

DESIGN ISSUE #1:

STANDARD CONSIDERATIONS

- ▶ QofL measurements are particularly susceptible, more than physical measurements, to systematic errors associated with “observer effects” and the conditions under which the measurement is made.
- ▶ Standard precautions for avoiding bias should be adopted: randomization, blindness of the questionnaire's administration, as well as the standardization of recording procedures.

DESIGN ISSUE #2:

CHOICE OF QofL INSTRUMENT

- ▶ A key choice to make is between the use of a standard instrument and a specifically developed questionnaire.
- ▶ Standard questionnaire may lack questions or sensitivity compared to a specifically developed questionnaire.
- ▶ However, standard questionnaires have a history of successful use; their validity and reliability may have been extensively tested - if in the same area/field .
- ▶ In general, experts recommend the use of a standard, validated instrument; if it is felt that “supplementary questions” are appropriate and desirable, they could be formed and added to the “core” instrument.

EXAM #3.2:

ARTHRITIS & RHEUMATISM

Rheumatoid arthritis (RA) assessment has special needs; several RA-specific instruments have been developed with arthritis impact measurement scales. However, generic health status instruments have also quite widely used as outcome measures in RA. For example the “Sickness Impact Profile” (SIP) is a very popular multi-dimensional instrument which has been extensively tested for general use (like SF36). It has been found to have higher test-retest reliability than RA-specific instruments, and to measure meaningful clinical changes produced by interventions.

DESIGN ISSUE #3:

CHOICE OF QoL DIMENSIONS

- ▶ What to measure in a clinical trial depends on the nature of the disease, expected benefits, adverse effects, and the length of the observation period.
- ▶ For example, in severe illness, patients are unlikely to be working or being physical active; questions on these dimensions would not be included.
- ▶ Patients should not be burdened with inappropriate questions and to an approach that is insensitive to small but important changes.

EXAM #1.2:

TREATMENTS FOR NEUROBLASTOMA

The QofL questionnaire for the neuroblastoma trial addressed both physical and psychological aspects of the child's QofL. Topics included "functional status" (restriction of physical activity), symptoms (including pain), side-effects (nausea and vomiting, loss of appetite, difficulties in hearing), and worry about side-effects (such as hair loss), and overall assessment of enjoyment of life. An inclusion of additional items on social functioning of the child, and on the impact of the disease and treatment on the family and friends of the child, was judged as undesirable.

DESIGN ISSUE #4:

CHOICE OF SCALE

There are several typical ways of recording answers

- (1) As binary response, e.g. condition present or absent
 - (2) As a response on a k-point ordinal scale in increasing or decreasing severity, e.g. $k = 5$.
 - (3) On a visual analog scale, by marking a point on the line on representing increasing severity from the left.
- ▶ An advantage of the visual analog and ordinal scales is the protection against loss of information; however, visual analog scale involves more work on analysis - even it offers more possibility in post hoc grouping.
 - ▶ There is no gain in going beyond 5-point scale (but there still are instruments with 7-point scale!)

DESIGN ISSUE #5:

WHO AND WHOM TO MEASURE

The essence of the QofL approach is the expression of a “subjective viewpoint”, not to avoid it; therefore the main respondent, whenever feasible, should always be the patient. Of course, inability of a patient to respond adequately (e.g. mentally disabled or severely ill patients) may necessitate “proxy assessment” by a relative or professional. Consistency in the later case should be emphasized. For example, the median age of children in the “Neuroblastoma Trial” was about 3 years, proxy assessment and reporting of QofL was generally necessary. However, agreement between parental and clinician assessments have been found to be rather poor.

DESIGN ISSUE #6:

WHERE AND WHEN TO MEASURE

- ▶ Sometimes, it might be acceptable as well as convenient to administer QofL surveys by phone; however, it may lead to loss of sensitivity.
- ▶ When to measure QofL is largely dictated by the objectives of the trial. Generally, “statistical principles” apply: (i) the need for “base-line” data, and (ii) assessment should be specific to some well-defined period (e.g. the last few days or last month), and (iii) there should be a final assessment of QofLife in patients who withdraw from the follow-up.
- ▶ QofL changes may not be apparent if the follow-up period is too short since patients may take time to modify a life style adapted to a chronic condition.

Design versus Data Analysis

- ▶ QofL concept started as “global & holistic”; a more reductionist approach (in implementation) is necessary.
- ▶ The concept is broken into components called dimensions; e.g. “Physical well-being”, “Psychological well-being, and “Social well-being”.
- ▶ Each dimension is further decomposed into a number of questions called ‘items’; which in turn may be answered on a “scale” - for example, 5-point scale, 10-point scale.
- ▶ In describing and comparing groups, or “arms” of a clinical trial, one “combines” the detailed responses back into dimensions - or into a single “global assessment”.

Design: “Decomposition” into “components”

Analysis: “Re-composition” from components

Steps in the design stage are less controversial. The reverse process in the “data analysis stage”, the process of “re-composition”, from items into dimensions or into a global “score” involves the use of weighting schemes - which may provokes further concerns and critiques.

HOW WEIGHTS ARE ASSIGNED?

- ▶ The need for some “recombination” from items’ scores requires a set of weights which implicitly “quantifies” different states of health.
- ▶ The effect of the differential weights can be very severe; decisions might be different for different sets of weights!
- ▶ There are two main approaches to assign weights: (1) use methods such as “Factor Analysis” (a statistical method) to identify “constructs that underlie the responses, or (2) to scale the states according to implicit or explicit personal valuations.

IMPROVEMENTS?

- ▶ Can we get out of the mess? No and Yes!
- ▶ Cox et al (1992), a very famous statistician, suggested that, when it involves mortality, “it is pointless to conduct elaborate experiment to derive sophisticated weighting schemes”; no weights can be completely and satisfactorily justified, one needs to analysis/compare quality of life data conditional on survival.

GUIDELINES ON WEIGHTING

For instrument adoption, formal psychometric methods may be needed but, for easy interpretation of outcomes, simple choice of weights are suggested: (1) simple scores for answers to items - such as 0,1,2,...or express as percent of maximum achievable score; (2) unweighted average over item answers expressing in above standardized form; (3) dimension means, dimension-by-treatment means are key summaries; (4) If needed, unweighted average over dimensions in the absence of treatment-by-dimension interaction; (5) If necessary, to test the significance of treatment difference, use Bonferroni-type adjustment to account for multiple decision (one for each dimension); (6) The restriction of analyses to a few items nominated “a priori” also helps to avoid sticky issues.

ON GLOBAL DECISION

If needed, unweighted average over dimensions in the absence of treatment-by-dimension interaction. Secondly, if it is likely that the treatments affect all dimensions similarly, it may be sensitive (most early-phase trials are small) to put more emphasis on “overall means” (across dimensions). However, for some items, e.g. adverse affects, it may be wise to deal separately to distinguish frequency from severity.

In summary, special aspects of the analysis of QofL data in a clinical trial setting stem partly from the essentially multi-dimensional characteristic of the concept and partly from the substantial components of variation not only between patients within treatment groups but also across time within patients. There are often further complications because of patient withdrawal and a need to consider QofL data alongside information on survival and other clinical outcomes.

NATURAL HISTORY OF DESIGNS

1) More QofL research are conducted on “prostate cancer” than any other cancers or other diseases because of an area not covered adequately by clinical outcomes: possible “sexual dysfunction”, including radiation-induced impotence, which may negatively affect patients’ relationships - as well as their own self-esteem.

2) Interest in QofL assessment for patients affected by prostate cancer has been on the rise because of another specific reason besides the effects on patients’ social and vocational activities. It’s the slow progress of the disease; untreated patients may live for many years after diagnosis, then may die from other pathologies. The debate is whether to treat; and Quality of Life might be a logical criterion for making that crucial decision?

Example #4A:

PROSTATE CANCER FOLLOWING
RADIOTHERAPY

Roach et al (1996); Int. J Radiation Oncology &
Biol. Physics

Example #4B:

PROSTATE CANCER FOLLOWING
RADIOTHERAPY

Caffo et al (1996) on British Journal of Urology

Both early studies, 4A and 4B, represent the most simple both in the Design and the Analysis for Quality of Life studies:

1) self-made questionnaire, questions may be culled from various sources or made up by investigators; if so, some psychometric evaluation is necessary (not done in 1A!). 2) Questionnaire is administered once, usually by mail -after the completion of the treatment, good participation, 3) Demographic factors and others (e.g. stage) are assessed by Chi-square, t-test/Wilcoxon, or one-way ANOVA, 4) Descriptive results proportion, mean, maybe correlation 5) The only comparisons made are “before” versus “after treatment”; done by standard methods: McNemar Chi-square, one-sample t-test/Signed-rank Wilcoxon etc...

Example #5A:

METASTATIC NON-SMALL CELL LUNG CANCER

Finkelstein et al (1988) on American Journal of Oncology

Example #5B:

**RADICAL PROSTATECTOMY VERSUS RADIATION FOR
PROSTATE CANCER**

Lim et al (1995) on The Journal of Urology

QofL INSTRUMENTS USED

- 1) FLIC: 22 items, each rated on 1-7;
- 2) “The Profile of Mood States” (PMS) has 65 items in 6 dimensions, each is rated on 5-point scale from 0 (not at all) to 4 (extremely).
- 3) A “Symptom Inventory” questionnaire was constructed to evaluate bladder irritability, symptoms of urinary incontinence, sexual dysfunction, and bowel dysfunction. Each dimension has 5-10 items - rated on 5-point scale from 0 (not at all) to 4 (extremely).

More Complicated Designs:

Study 5A extended the design of examples 4A & 4B in one direction: measurements repeated over time;

Study 5B extended the design of examples 4A & 4B in another direction: from 1 to 2 groups;

Example #6A:

PHASE II STUDY OF FLUOROURACIL AND ITS
MODULATION IN ADVANCED COLORECTAL CANCER

Leichman et al (2000) on Journal of Clinical Oncology

Example #6B:

CHALLENGES BY NON-RANDOM QUALITY OF LIFE
MISSING DATA IN AN ADVANCED-STAGE COLORECTAL
CANCER CLINICAL TRIAL

Moinpour et al (2000) on Psycho-Oncology

THE NEXT STEP IN DESIGNS

- ▶ These studies, 6A and 6B, are extensions in both: multi-arm trial where QofL assessment is repeated over time. Use “ANOVA for repeated measurements” (RM).
- ▶ Consider to take a course in “Longitudinal Data” or “Correlated Data”; showing here is the ANOVA Table and the tests.

Repeated Measures ANOVA

Abbreviated ANOVA Table:

Source	SS	df
Group	SS_G	$k-1$
Subjects (Group)	$SS_{S(G)}$	$n-k$
Time	SS_T	$t-1$
Group x Time	SS_{GT}	$(k-1)(t-1)$
Residual	SS_R	$(n-k)(t-1)$

DIFFERENCES AMONG GROUPS

An F-test with $(k-1, n-k)$ degrees of freedom:

$$\begin{aligned} F &= \frac{MS_G}{MS_{S(G)}} \\ &= \frac{SS_G / (k - 1)}{SS_{S(G)} / (n - k)} \end{aligned}$$

TIME EFFECTS

An F-test with $[t-1, (n-k)(t-1)]$ degrees of freedom:

$$\begin{aligned} F &= \frac{MS_T}{MS_R} \\ &= \frac{SS_T / (t - 1)}{SS_R / (n - k)(t - 1)} \end{aligned}$$


GROUP-BY-TIME INTERACTION

An F-test with $[(k-1)(t-1), (n-k)(t-1)]$ degrees of freedom:

$$\begin{aligned} F &= \frac{MS_{GT}}{MS_R} \\ &= \frac{SS_{GT}/(k-1)(t-1)}{SS_R/(n-k)(t-1)} \end{aligned}$$

A DIFFICULT AREA

- ▶ Missing data may not be missing by chance because they are associated with factors that are also associated with poor study outcomes.
- ▶ It has been suggested that it would be biased (i) to compare either all patients at baseline to all patients to fill out forms at each time, or (ii) to use only patients with baseline and all subsequent forms
- ▶ In (i), if patients with seriously decreased quality of life at time t do not fill out forms, the average at that time is biased toward good QofL; in (ii) the difference among patients who fill out all forms may not be representative of the whole study sample



It has been suggested by some investigators that one may consider a series of analyses: all patients with baseline and first assessments, all patients with baseline and first two assessments, all patients with baseline and first three assessments, etc... The results would tell if the reason for discontinuing QofL assessments is due to illness or death.

PRAGMATIC DECISION

- ▶ Clinical trials often used a variety of endpoints to evaluate the costs and benefits of alternative treatments.
- ▶ Some may argue that separate analyses of these endpoints are conventional and scientifically valid, but may be inadequate for a practical, overall decision on which treatment is most likely to benefit a “particular” patient.
- ▶ If all relevant endpoints favor one treatment, the choice is clear; but if one shows an advantage on some endpoint but a disadvantage on another, the conclusion is more difficult. For example, the more toxic therapy may also result in improved response and overall survival.

TRADE-OFFS

- ▶ When the results are different for different endpoints, the recommendation will depend on two factors: (i) the size of the difference and (ii) the “relative importance” of each outcome. The latter is more subjective.
- ▶ With separate analyses of different endpoints, there is a danger that statistical significance would be given undue emphasis over clinical significance; for example, a month more of overall survival versus more severe toxicity.
- ▶ Of course, there is always the difficulty of combining individual outcomes because such a combination unavoidably depends on the “weighting” values chosen.

Suggested Readings:

Search, find (& Read) the papers in the following examples

Example #5A:

METASTATIC NON-SMALL CELL LUNG CANCER

Finkelstein et al (1988) on American Journal of Oncology

Example #5B:

RADICAL PROSTATECTOMY VERSUS RADIATION FOR
PROSTATE CANCER

Lim et al (1995) on The Journal of Urology