June, 1997

Effect size and power for clinical trials using years of healthy life as the primary endpoint

Paula Diehr, University of Washington

Available at: https://works.bepress.com/paula_diehr/29/
EFFECT SIZE AND POWER FOR CLINICAL TRIALS THAT MEASURE YEARS OF HEALTHY LIFE

PAULA DIEHR, BRUCE M. PSATY AND DONALD L. PATRICK

SUMMARY

Some clinical trials perform repeated measurements on patients over time, plot those measures against time, and summarize the results in terms of the area under the curve. If the measured variable is health status, the summary outcome is sometimes referred to as years of healthy life (YHL), or quality-adjusted life years (QALY). This paper investigates some theoretical and practical aspects of randomized trials designed to assess measures such as YHL. We first derived algebraic expressions for the effect size of YHL measures under several theoretical models of the treatment's effect on health. We used these expressions to examine how the length of the study, the number of measurements per person and the correlations among health measurements over time influence the effect size. We also explored the relative statistical power of analyses based on YHL versus analyses based on change-scores using the same data. We present an example. Findings suggest that: (i) the number of measurements per person need not be large; (ii) high correlation among measures over time tends to lower the power of a study using YHL; (iii) a longer study will not always provide more power than a shorter study, and (iv) analyses based on YHL may have less power than change-score analyses. Some of these findings depend on the model of change in health status caused by the treatment. Such models require further study. © 1997 by John Wiley & Sons, Ltd.

1. INTRODUCTION

Clinical trials increasingly include sequential measures of health-related quality of life, in addition to the more traditional assessments of morbidity and mortality. These health status measurements are often summarized by plotting the measurements over time and calculating the area under the curve. If the health status instrument was developed with appropriate weights, one can interpret the area under the curve as years of healthy life (YHL), a measure of health outcome that integrates quality of life and duration of survival. This quantity is also known as quality-adjusted life years (QALY).

Because YHL-like measures provide an estimate of ‘effectiveness’ for cost-effectiveness analyses, they have been used with increasing frequency in clinical trials and programme evaluations that measure both the cost and outcomes of treatments. For example, a randomized trial...
evaluated the cost and effectiveness of providing preventive services to Medicare patients. Investigators hypothesized that the introduction and reimbursement of these preventive services would improve quality of life and increase the years of healthy life in the treatment group. One goal of the study was to determine the incremental cost per year of healthy life gained, by measuring health status at baseline, year 2, and year 4, and using these measurements to estimate the area under the curve for the treatment and control groups. The difference in area under the curve between the treatment and control groups is the years of healthy life gained attributed to the intervention. The cost of the preventive services packages was also measured.

The Quality of Well-Being Scale (QWB) was used in this study. On the QWB, health status is scaled between one (perfect health) and zero (death), using weights derived through preference-rating, so that the area under the curve has meaning. Other appropriate instruments are also available. One can also construct some simple measures from responses to a question such as ‘was the person symptom-free’, where yes and no = 0. The average of these binary responses is the proportion of people symptom-free, and the area under such a curve is an estimate of the average years of symptom-free life. ‘Area under the curve’ (AUC) measures are of interest in other studies such as pharmacokinetics trials that compare different formulations of the same drug and asthma trials that use repeated clinical measurements such as FEV measurements.

Although there is some literature about the design of such trials, there are statistical issues not yet addressed. This paper explores the factors most important in the design of an efficient clinical trial with years of healthy life as a primary endpoint. The findings are also relevant to other studies that involve the area under a curve.

METHODS

We first develop notation for a ‘typical’ study that measures health status over time, and then propose three simple models of the types of change in health that an intervention might cause. For these models, we derive algebraic expressions for the area under the curve (years of healthy life), and for the effect size of YHL measures. We use these equations to examine the influence on effect size of the length of the study, the number of measurements per person, and the correlations among the health status measures over time. We also compute the effect size for analyses based on change-scores, for each model, and compare the power of a change-score analysis to the power of a YHL analysis. We also present an example drawn from the study of health promotion in older adults.

FINDINGS

Estimates of YHL in a randomized clinical trial

In a typical study, one measures health status at baseline and then ‘K’ times thereafter, usually at regular intervals, until the end of the study. (In a simple pre-test/post-test study, K = 1). In our notation, at times \( t_0, t_1, \ldots, t_K \), during the study one measures a person’s health score, \( h_0, h_1, \ldots, h_K \). For simplicity, we let \( t_0 = 0 \), and assume that the measurements occur at regular intervals. The total length of the study is \( t_K \). Figure 1 illustrates the design.

We can use the study data to estimate the area under the curve, or years of healthy life. The simplest estimate of the area under the curve comes from a trapezoidal or ‘connect the dots’ strategy, with

\[
\hat{YHL} = \hat{\text{AREA}} = \frac{1}{2} (h_0 + h_K + 2 \sum_{j=1}^{K-1} h_j) t_K. \tag{1}
\]
For example, in a pre-test/post-test design that lasts 2 years ($K = 1$ and $t_k = 2$), $\hat{A}$ reduces to $h_0 + h_1$.

The variance of $\hat{A}$ is a complicated function of the variances and intercorrelations of the $h_i$, as shown in the Appendix. To simplify the form of the estimate, we next consider the special case where the variance of the health measure is identical at each time point, and the correlation between any two health measures for an individual is the same, no matter how far apart in time the measures are taken. Later, we address the appropriateness of these assumptions in the example and in the discussion. Under these simplifying assumptions, the variance of the estimate of YHL for one person is

$$\text{var}(\hat{A}) = \frac{\sigma^2[(2K - 1) + \rho(2K^2 - 2K + 1)]t_k^2}{2 K^2}$$

(2)

where $\sigma^2$ is the variance among subjects in measured health status at a particular time. The term $\rho$ is the correlation between a person’s measured health at, say, $t_0$ and $t_1$. If the person’s true health has not changed between the two times, then the measured health differs only because of measurement error. In that case, $\rho$ is known as the reliability of the health status instrument. The variance of the estimated area is proportional to $\sigma^2$ and to $t_k^2$, and also increases with $\rho$.

**Theoretical models of change**

We next pose three theoretical models of how average health status might change as the result of a treatment intervention, and derive algebraic expressions for the effect size of YHL measures under each model. Figure 2 shows the models.
In all three models, the control group has mean health status $h_0$ at time $t_0$, which decreases to health 0 (death) at time ‘$a$’ (the heavy solid line). In model 1 (different slopes), the treatment group also begins with mean health $h_0$ (because of randomization), and decreases in health status, but the decline is not as rapid as that in the control group. The treatment group reaches zero health at time ‘$b$’ (M I, the dashed line in Figure 2). Model 1 might be appropriate for a persistent behaviour change such as increase in exercise or reduction in cholesterol.

In model 2 (parallel lines), the treatment takes effect almost immediately, and rapidly achieves a maximum health effect that is maintained until death at time $b$ (M II). Model 2 might describe the change in health status for some acute but long-lasting intervention such as arthroscopic surgery of the knee to remove torn cartilage. In model 3, the treatment group registers a short-term improvement, after which the treatment and control subjects have the same health (M III). A possible example of model 3 is an intervention such as a ‘crash’ diet after which the subject regains the lost weight. There are, of course, many other possible models of change in health status. We choose these three because they are simple to present, and have some interesting features.

For these models, we can compute the theoretical area under the curve and the resulting treatment differences using basic geometry. Let the time of the baseline measurement, $t_0$, be zero. Under model 1 the years of healthy life for the control group is the area under the triangle described by $(h_0, t_0, a)$, which is one-half the base times the altitude, or $A_C = (a/2)h_0$. Similarly we can calculate the area under the curve for the treatment group, $A_T$. The difference between the treatment and control YHL is the treatment difference,

$$A_T(1) - A_C(1) = h_0(b - a)/2.$$
Although we may expect many interventions to confer benefits throughout life, it is unusual for a study to follow all subjects until death. With subjects followed only until time $t_K$, the difference in YHL for the two groups up to time $t_K$ is

$$A_{T(1)} - A_{C(1)} = h_0 t_K^2 (b - a)/(2ab). \quad (3)$$

Under model 2, assuming that the treatment takes effect very close to $t_0$, similar considerations yield a treatment difference measured only up to time $t_K$ of

$$A_{T(2)} - A_{C(2)} = h_0 t_K (b - a)/b. \quad (4)$$

Under model 3, the treatment difference is the area of the small sector where the treatment and control health diverge. This area, which we call $Q$, is unrelated to the length of the study, $t_K$, as long as the effect has disappeared by $t_K$, as assumed here.

**Effect size for different designs**

We next consider the effect size of a treatment under the three hypothetical models. The term ‘effect size’ (ES) is used in many ways.\(^{20,21}\) Here, we define ES as the true treatment difference ($\Delta$) divided by the standard deviation of the estimated treatment difference in a hypothetical experiment with one person in each treatment group, which is the square root of ($\sigma_1^2 + \sigma_2^2$). Under this definition, $ES^2 = \Delta^2/(\sigma_1^2 + \sigma_2^2)$, a term in the familiar sample size calculation for the number of subjects per group needed to detect a difference of size $\Delta$ between two group means with power $1 - \beta$:

$$N = \frac{(Z_{1 - \eta/2} + Z_{1 - \rho})^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2} = \frac{(Z_{1 - \eta} + Z_{1 - \rho})^2}{ES^2}. \quad (5)$$

If ES is large, the desired power can be achieved with a relatively small sample size. A large effect size thus permits a smaller $N$ for the same power, or more power for the same sample size. In our notation, the effect size for YHL is:

$$ES = \frac{(A_T - A_C)}{SD(A_T - A_C)}.$$

For model 1, equation (3) shows the true treatment difference, and, assuming the variance is the same in the two groups, the variance of $A_T - A_C$, (for one person per group) is twice the variance shown in equation (2). The effect size for a YHL analysis when model 1 is true becomes:

$$ES_{YHL(1)} = \frac{h_0(b - a)}{2ab \sigma} t_K \frac{K}{\sqrt{[2K - 1 + \rho(2K^2 - 2K + 1)]}}. \quad (6)$$

The ES for model 1 thus factors into three terms. The first term contains parameters determined by the health status instrument used, the initial health of the subjects, and how well the treatment succeeds in modifying health. The second term is $t_K$, the length of the study. Effect size increases linearly with the length of the study.

The third term in equation (6) is a function of $\rho$ and $K$. Table I shows some values of the third term for model 1, for various values of $\rho$ and $K$. (The tabled quantity is equivalent to the effect size if the product of the first two terms happens to equal 1). Note that the effect size gets smaller for larger values of $\rho$, and that this effect is stronger when $K$ is large. Similarly, increasing the number of follow-up measures, $K$, improves the effect size if $\rho = 0$, but makes little improvement for larger values of $\rho$. 

Table I. Effect size* of YHL for selected values of $\rho$ and $K$ under model 1

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0·00</td>
<td>1·00</td>
<td>1·15</td>
<td>1·34</td>
<td>1·51</td>
<td>1·68</td>
</tr>
<tr>
<td>0·50</td>
<td>0·82</td>
<td>0·85</td>
<td>0·88</td>
<td>0·91</td>
<td>0·92</td>
</tr>
<tr>
<td>1·00</td>
<td>0·71</td>
<td>0·71</td>
<td>0·71</td>
<td>0·71</td>
<td>0·71</td>
</tr>
</tbody>
</table>

*The effect size under model 1 if terms 1 and 2 in equation (16) are set to 1. This term also occurs in the effect size equations for model 2 and model 3.

For model 2, under the assumption that the treatment takes effect almost instantly, the effect size is

$$ES_{YHL(2)} = \frac{h_0(b - a)}{a\sigma} \frac{K}{\sqrt{[2K - 1 + \rho(2K^2 - 2K + 1)]}}.$$  

Notably, this effect size is independent of $t_K$; that is, under model 2, a longer study provides no more power than a shorter study with the same number of measurements. The relative effects of $K$ and $\rho$ are the same as in model 1, and are shown in Table I.

Under model 3, the true difference in YHL is a constant, say $Q$. The effect size is then:

$$ES_{YHL(3)} = \frac{Q}{\sigma} \frac{K}{t_K \sqrt{[2K - 1 + \rho(2K^2 - 2K + 1)]}}.$$  

In this situation, $t_K$ is in the denominator, and the effect size actually decreases as the length of the study increases. The relative effects of $K$ and $\rho$ are the same as in model 1.

**Change-scores versus YHL**

It is common, in pre-test/post-test studies, to use the change-score, $CS = h_0 - h_K$, as the dependent variable. We next consider whether a study with the change-score as the endpoint is more or less powerful than a study using YHL as the endpoint. Under mode 1, the treatment difference in change-scores is

$$A_T - A_C = h_0 t_K (b - a)/(ab).$$

This is true for all three models. For model 3, however, $a = b$, because the treatment difference disappears. The true treatment difference for a change-score is thus zero.

In estimating the change-score, for one person,

$$CS = h_0 - h_K$$

and

$$\text{var}(CS) = 2\sigma^2 (1 - \rho).$$

The effect size for the change-score is then

$$ES_{CS} = \frac{h_0(b - a)}{2a b\sigma} \frac{t_K}{\sqrt{(1 - \rho)}}.$$  

(7)
For YHL under model 1, using equation (6), setting \( K = 1 \)

\[
ES_{YHL(1)} = \frac{h_0(b - a)}{2a b \sigma} \frac{t_K}{\sqrt{(1 + \rho)}}.
\]  

The effect sizes in equations (7) and (8) differ only by the sign of \( \rho \) in the denominator. \( ES_{CS} \) increases as \( \rho \) increases, but \( ES_{YHL(1)} \) decreases as \( \rho \) increases. Since \( \rho \) is almost sure to be greater than 0, the effect size for the change-score analysis is always larger than the effect size for the analysis based on YHL. This finding means that, under model 1, a study based on YHL necessarily has less power than a study using the same data but based on change-scores. One can somewhat improve the power of YHL by taking more than one post-baseline measurement \((K' > 1)\).

For model 2, parallel lines, the effect size for the change-score is as in (7), but the effect size for YHL (again, for \( K = 1 \)) is

\[
ES_{YHL(2)} = \frac{h_0(b - a)}{a \sigma} \frac{1}{\sqrt{(1 + \rho)}}
\]

and the ratio of the effect sizes is

\[
\frac{ES_{YHL(2)}}{ES_{CS}} = \frac{2b}{t_K} \frac{\sqrt{(1 - \rho)}}{\sqrt{(1 + \rho)}}.
\]

Under model 2, if \( \rho \leq 0.6 \), then YHL is more powerful than CS, since \( t_K < b \). If \( \rho > 0.6 \), YHL is also more powerful than CS if \( t_K \) is substantially less than \( b \). For example, if \( \rho = 0.8 \), then YHL has a higher effect size than CS if \( t_K/b < 2/3 \); that is, if the study takes up less than 2/3 of the average remaining years of life, which is usually the case. The YHL analysis is thus usually more powerful than the change score analysis.

For model 3, the short-term effect, the true treatment difference for the change-score analysis is zero. That is, for a change-score analysis, the null hypothesis is true, since the change at \( t_0 \) was identical in the two groups. There is, however, a true difference, \( Q \), in years of healthy life, which we might have detected by the YHL analysis if we had measured some points while the treatment group was doing better. Under the design shown in Figure 2, the YHL analysis would also have shown no treatment difference for \( K = 1 \), even though there was a difference. For this reason neither method is appropriate for \( K = 1 \), but the YHL method is appropriate for \( K > 1 \) if one makes at least some of the observations while the treatment effect is present.

**Adjustment for baseline covariates**

It may be advisable to adjust the YHL for some baseline covariates, to improve power. Consider the simple adjustment in which each \( h_{ijk} \) is replaced by \( h'_{ijk} = (h_{ijk} - h_{ij0} + h_{i.-0}) \), where \( i, j, \) and \( k \) index, respectively, the treatment group, the person within treatment, and the time of the measurement. That is, subtract each person’s first measurement from the other measures, and add in the treatment group mean at \( t_0 \). One can estimate the adjusted area under the curve by substituting \( h' \) for \( h \) in equation (1). The variance of the adjusted area, under the previous assumptions, is

\[
\text{var} (\hat{A}_{\text{adjusted}}) = (2K^2 - 1) \sigma^2 (1 - \rho) \frac{t_K^2}{2K^2}.
\]

For adjusted YHL, higher correlation among measurements thus yields a smaller, rather than a larger, variance. The variance of the unadjusted estimate in equation (3) is larger than the
Table II. Summary statistics from example

<table>
<thead>
<tr>
<th>Time</th>
<th>$t_0$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_0$</th>
<th>$t_1$</th>
<th>$t_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean QWB ($h_0$)</td>
<td>0.70</td>
<td>0.68</td>
<td>0.63</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>SD</td>
<td>0.10</td>
<td>0.16</td>
<td>0.21</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Correlation of $h_i$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with $h_0$</td>
<td>1.00</td>
<td>0.36</td>
<td>0.27</td>
<td>1.00</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>with $h_1$</td>
<td>0.36</td>
<td>1.00</td>
<td>0.63</td>
<td>0.45</td>
<td>1.00</td>
<td>0.48</td>
</tr>
<tr>
<td>with $h_2$</td>
<td>0.27</td>
<td>0.63</td>
<td>1.00</td>
<td>0.39</td>
<td>0.48</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Least squares equations:
- Linear $h = 0.70 - 0.0175t$
- Quadratic $h = 0.70 - 0.0025t - 0.00375t^2$

Estimated YHL:
- Trapezoid YHL | SD (YHL) | YHL | SD (YHL) |
  | 0, 1, 2 | 2.69 | 0.52 | 2.81 | 0.31 |
  | 0, 2 | 2.68 | 0.51 | 2.80 | 0.32 |
- Adjusted YHL | SD (YHL) |
  | 0, 1, 2 | 2.69 | 0.46 | 2.79 | 0.27 |
  | Integrate quadratic | 2.72 |

adjusted variance if $\rho > (K - 1)/(2K - 1)$. If $K = 1$, the adjusted estimate has lower variance than the unadjusted variance for any positive value of $\rho$. In fact, for $K = 1$, adjusted YHL is equivalent to the change score. For $K = 2, 3$ and $4$, the respective threshold values of $\rho$ are 0.33, 0.40 and 0.43. For large values of $K$, the threshold value approaches 0.50. For high values of $\rho$, adjustment for baseline covariates should be considered.

**Example**

In the preventive services evaluation described in the introduction, the QWB was measured for 2279 older adults at baseline, and again at 2 and 4 years after baseline ($K = 2$, $t_K = 4$). Unfortunately, the treatment group achieved fewer years of healthy life than the control group during the experiment. We have combined the two groups for this example. Table II shows summary measures for all subjects with complete data, and also for the 2088 subjects who survived to $t_2$.

The mean QWB at the three times was 0.70, 0.68 and 0.63, not quite a linear decline. The decline was due in part to subjects who died, who received QWB scores of zero. The survivors experienced less decline, from 0.71 to 0.69, as shown in the table. Linear and quadratic equations were fit to the means and are in the table. The standard deviations for all subjects increased over time, from 0.10 to 0.21, while the SDs for the subset of survivors remained stable. The increase over time for all subjects resulted from the people who died, since they received a score of zero, quite far from the mean. For this reason, the SD would probably continue to increase until about half of the subjects died, after which it would decrease. The correlations for all people were unstable, and became higher over time (for example, $\rho_{12} = 0.63$). Correlations for survivors were
Table III. Sample size calculations†

<table>
<thead>
<tr>
<th>$t_K$</th>
<th>$K$</th>
<th>2</th>
<th>4</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>$\rho$</td>
<td>$N$</td>
<td>$N$</td>
<td>$N$</td>
<td>$N$</td>
</tr>
<tr>
<td>0.1</td>
<td>0.27</td>
<td>3410</td>
<td>853</td>
<td>2695</td>
<td>674</td>
</tr>
<tr>
<td>0.1</td>
<td>0.63</td>
<td>1025*</td>
<td>4459*</td>
<td>1115*</td>
<td>4459*</td>
</tr>
<tr>
<td>0.22</td>
<td>0.27</td>
<td>16506</td>
<td>4126</td>
<td>13043</td>
<td>3260</td>
</tr>
<tr>
<td>0.22</td>
<td>0.63</td>
<td>23336*</td>
<td>5834*</td>
<td>21581*</td>
<td>5395*</td>
</tr>
<tr>
<td>0.16‡</td>
<td>0.42‡</td>
<td>10235*</td>
<td>2559*</td>
<td>8780</td>
<td>2195</td>
</tr>
<tr>
<td>0.16‡</td>
<td>0</td>
<td>6021*</td>
<td>1505</td>
<td>3512</td>
<td>878</td>
</tr>
</tbody>
</table>

*Adjusting YHL for baseline covariates would allow a smaller sample size
†Necessary sample size per group for 80 per cent power for a study where $h_0 = 0.70$, $a = 40$, $b = 56$, $t_K$, $\sigma$ and $\rho$ are as specified in the table, and model 1 is true.

Definition of terms: $t_K$, length of study; $K$, number of measurements after baseline; $\sigma$, standard deviation of health for one person; $\rho$, correlation between health at one time and health at another time
‡Average of values in Table II

fairly stable at about 0.4. The instability for all subjects also resulted from the people who died, since their scores after death were outliers, and were also perfectly correlated.

The estimated area under the QWB curve, using equation (1), was $YHL = 0.70 + 2(0.68) + 0.63 = 2.69$ years of healthy life out of a possible 4.0. Using only $h_0$ and $h_K$, $YHL = 2.66$. Integrating the quadratic equation in Table II yields an area of $YHL = 2.72$ years of healthy life. The standard deviations of the first two YHL estimates, shown in Table II, are also similar. The middle observation thus changed the estimated area in only trivial ways, and the method of calculating the area also did not matter much.

We noted above that when $K = 2$ a simple adjustment for baseline health status would reduce the variance of the estimated YHL if $\rho > 0.33$, which is true for this data set. The adjusted YHL estimates are shown in Table II, and do have lower variance than the unadjusted YHL values. For all cases and for only those alive at $t_2$, adjustment resulted in about a 22 per cent reduction in the variance of the YHL estimate, with a corresponding decrease in required sample size.

Planning a study

As an example, we next design a new study in which we expect the control group’s health to decrease from 0.70 to 0.63 in 4 years, as in Table II. The line that connects these points crosses the $x$-axis at $t = 40$ years that is, $a = 40$. Assume we know that model 1 (different slopes) holds and we hypothesize that in the treatment group health decreases less steeply, from 0.70 to 0.65 in 4 years. This decline is equivalent to $b = 56$ years.

We calculated the effect sizes from equation (6) and the sample sizes from equation (5) required for 80 per cent power, for several values of $\sigma$ and $\rho$ taken from Table II. The results appear in Table III. For example, with $\sigma = 0.1$, $\rho = 0.27$, $t_K = 2$, and $K = 2$, the required sample size is 3410 per group. Lengthening the study to 4 years decreases the sample size to 853 per group. Using four measurements in a two-year study requires 2695 per group, and four measurements in a four-year study requires 674 per group. One would probably prefer a four-year study with $K = 2$ follow-up measurements, since one would need fewer total surveys. Using higher values
of $\sigma$ and $\rho$ causes substantial increases in the necessary sample size, as shown on other lines of Table III. The asterisks denote situations in which power could be increased by using the adjusted value of YHL.

We might have used the quadratic equation of Table II instead of the assumption that health declined linearly over time. However, as noted above, the linear and quadratic estimates of YHL were quite close. Further, since the error caused by using an improper model will occur in both treatment and control subjects, the error is likely to cancel out when we subtract the treatment and control areas to calculate the treatment difference.

We also used the equation in the Appendix, which does not require $\rho$ and $\sigma$ to be time invariant. Consider a four-year study with two follow-ups after baseline ($t_K = 4$ and $K = 2$). Using the parameter estimates from Table II (for all), and the equation in the Appendix, the estimated variance of the estimated YHL is 0.2844. If instead we use the average of the values in Table II in equation (2) ($\rho = 0.42, \sigma = 0.16$), the estimated variance is 0.2611. The estimated effect sizes and sample sizes based on these variances thus differ by only about 4 per cent. It seems reasonable to use the simpler equations. It is, however, important to use the average values of $\rho$ and $\sigma$, rather than the initial values.

**DISCUSSION**

We have examined the effect size and power of clinical trials based on repeated measures of health status, as a function of $\rho$, $K$, $t_K$, the nature of the treatment effect over time, and whether the endpoint is years of healthy life or the change in health from $t_0$ to $t_K$.

Under all models, a higher value of $\rho$, the correlation of a person’s health status measurements over time, improves power for a change-score analysis, but decreases power for a YHL analysis. Correlations tend to be positive because a person who is healthier than average at one time tends to be relatively healthy at a later time. As shown in Table I, the influence of $\rho$ on the effect size for YHL is not strong unless $K$ is large. Choosing a less reliable instrument would probably decrease the value(s) of $\rho$, but would also create a comparable increase in $\sigma$, and result in a study with lower rather than higher power.

One could reduce the correlation to zero in group-randomization experiments such as community interventions, which enroll a large number of people, but for which one measures only a sample of the people at each time.\textsuperscript{22} If we draw a fresh cross-sectional sample for each evaluation of health status, rather than follow a cohort over time, the correlation over time is zero. Such an approach also requires incorporation into the later measurements of an estimate of the deaths in the community since baseline. Comparison of the last two lines in Table III shows that we need to survey 40 per cent to 60 per cent fewer subjects each time under the repeated cross-sectional design ($\rho = 0$) than under a cohort design. Ramsey \textit{et al.}\textsuperscript{19} proposed a related approach for observational data. As noted above, adjustment for baseline covariates may reduce variability among subjects, which will also reduce the negative effect of $\rho$.

Under all models, increasing $K$, the number of follow-up measurements per person has no effect on the change-score analysis, which uses only the first and last measurement. Increasing $K$ also achieves surprisingly little improvement in power for the YHL analysis. This finding may seem counter-intuitive, since one would require many closely-spaced health measurements to be able to approximate the time pattern of health for an individual. In a randomized trial, however, we have interest in estimating the average trend over time, which is, in our models, linear. Since two points determine a straight line, it is acceptable to design a study with only a single follow-up if we are certain that the linear model is appropriate. Even if the relationship is non-linear, the average curve is likely to be fairly smooth, and will not require many observations to estimate it.
This result suggests that efficient designs that use YHL-like outcomes will include as few measurements as one can justify. Under model 3, however, a study with too few measurements could well miss the transient but potentially important treatment difference. It is reasonable that we obtain one measurement fairly early in the study, to make sure that there is at least one outcome point and that we do not miss early treatment differences. One final measurement is needed to estimate the long-term treatment difference. Unfortunately, with measurements at only three time points, we are not likely to learn the shape of the curve for use in future studies. If we are interested in describing the exact shape of the curve, additional measures per person are needed.

The effect of the length of the study, $t_K$, depends on the model of change in health. We should consider longer studies if we expect that model 1 holds. Doubling the length of time for a study (keeping other factors the same) allows us to perform the study with only 25 per cent as many subjects as the original study would have required for the same power. This increase in power also holds true for a change-score analysis under model 1. There may also be reasons not to perform longer studies. If $\sigma$ and $\rho$ increase over time, as is likely, the increase in power due to a longer study attenuates. A longer study costs more than a shorter study, and experiences more loss to follow-up, which compromises the internal validity of the study. Finally, if the treatment effect resembles model 2 or model 3, then the extra time and expense would not provide any additional power. Information about the model of change in health is crucial in the choice of the length of the study.

The choice of a YHL analysis versus a change-score analysis also depends on the model. Under model 1, a change-score analysis is most powerful. Under model 2, a YHL analysis is usually more powerful. Under model 3, the null hypothesis is actually true for the change-score but false for YHL. The conclusions regarding model 3 parallel the documented loss of power over time found in survival trials whose hazard ratios diminish over time. Although they are superficially similar, the change-score and YHL do not measure the same quantity. In particular, if $\sigma$ does not change over time, the YHL and change-score estimates are completely uncorrelated.

Another difference between the two measures is that the change-score ‘removes’ initial differences between groups, since each person is her own control. A YHL analysis perpetuates initial differences, and so requires some adjustment in non-experimental situations (under the simple adjustment proposed above, adjusted YHL and CS become identical).

These important differences mean that investigators need to think carefully about what model they expect to hold, and about which aspects of the treatment effect on health they wish to measure. In a short-term clinical trial with few deaths, both measures may be conceptually appropriate. For a trial with many deaths, a change-score might miss important improvements in quality of life that occurred in the treatment group before death; in this situation, YHL is conceptually superior. Cost-effectiveness analyses are better suited to YHL than to change-score analyses.

**Limitations of the research**

There are several limitations of the research. One is that the work assumes complete data, collected at the same times (relative to baseline) for all subjects. If there are substantial missing data, or if one does not make measurements on the same schedule for all subjects, or if one follows subjects for different lengths of time, this ‘subject-based’ approach may be inappropriate.

The formulae in the text use the simplifying assumptions that $\rho$ and $\sigma$ are constant over time. These assumptions were not supported in the example presented above, or in other data that we have examined where there were deaths. The more complex equations are in the Appendix. The text equations are approximately correct if we use the average values (rather than the baseline values) for $\rho$ and $\sigma$. The example supports such an approach.
The research also assumed that we can consider $\rho$, $t_k$, and $K$ independently. It is likely that changing one design factor alters the other as well. An investigator should use estimates from studies similar in design to the planned study.

We examined only three models of change, all linear. We chose these because they illustrated some interesting points, but they may not represent all situations of interest. Since most clinical trials last at most for a few years, it is likely that straight lines provide an adequate model over such a short period of time. To prove otherwise would require multiple measures per subject. We do not know of any literature on the relative shape of such curves for treatment and control groups, but one can study other models of interest in this same way. One must know, hypothesize, or assume the true model in order to calculate a particular effect size. The general findings about the effects of $\rho$ and $K$, however, are the same for all models, since they depend on the data, not on the theoretical model. The effect of study length is highly model dependent, but depends only on whether $A_T - A_C$ increases faster than, proportional to, or slower than $t_k$. One may find it easier to specify this relationship than to predict the entire model. Thus, even without knowledge of the true model, the general findings are useful.

The trapezoidal method of estimating the area under the curve is unsophisticated. Although there is no bias for the observed data points (assuming no losses to follow-up), the method used to interpolate may cause a biased estimate of the area. We used other approaches in the example, and they made little difference. More complex models require more observation periods, and will probably not yield more powerful studies. Clearly, we need more information on the trend of health over time for subgroups of interest.

We analysed years of healthy life only until the end of the study, rather than until all subjects had died. If one expects the intervention to confer life-long benefits, one needs to estimate the additional costs and benefits in some way. The effects of such estimation on power and effect size are beyond the scope of the current paper. Costs and benefits are often discounted, which makes later costs and benefits less important than those actually measured in the study. In this case, it may be adequate to plan a study based only on the effect size within the study period.

**Conclusion**

Although there are limitations, the work presented here should have use in the planning of clinical trials that use years of healthy life as an endpoint. Some of these results are independent of the model of change, but exact power calculations depend on knowledge of the model, and require data of the type in Table II. The work is also useful for the other clinical trials where one can represent outcome as the area under a curve.

**APPENDIX**

The variance of $\hat{A}$ is a function of the variances and intercorrelations of the $h_i$, as follows. Letting $\sigma_i$ be the variance of $h_i$, and $\rho_{ij}$ be the correlation between $h_i$ and $h_j$,

\[
Z_1 = \left[ \sigma_0^2 + \sigma_K^2 + 4 \sum_{i=1}^{K-1} \sigma_i^2 + 2 \rho_{0K} \sigma_0 \sigma_K \right]
\]

\[
Z_2 = 4 \left[ \sum_{j=1}^{K-1} \rho_{0j} \sigma_0 \sigma_j + \sum_{i=1}^{K-1} \rho_{iK} \sigma_i \sigma_K + 2 \sum_{i=1}^{K-2} \sum_{j=i+1}^{K-1} \rho_{ij} \sigma_i \sigma_j \right]
\]

\[
\text{var}(\hat{A}) = \frac{t_k^2}{4K^2} [Z_1 + Z_2].
\]
ACKNOWLEDGEMENTS

This project was supported in part by the Health Care Financing Administration (HCFA) under Cooperative Agreement No. 95-C-99161. Dr. Psaty is a Merck/Society for Epidemiological Research Clinical Epidemiology Fellow.

REFERENCES