DESIGNING A COST-EFFECTIVE CLINICAL TRIAL

BY

JOHN C. HORNBERGER, BYRON WM. BROWN, JR.
and JERRY HALPERN

TECHNICAL REPORT NO. 172

JANUARY 1995

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANTS
5 R01 CA59039-20 AND 5 R01 CA55325

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Designing a Cost-Effective Clinical Trial

By

John C. Hornberger, Byron Wm. Brown, Jr.
and Jerry Halpern

Technical Report No. 172
January 1995

Prepared Under the Auspices
Of
Public Health Service Grants
5 R01 CA59039-20 and 5 R01 CA55325

Division of Biostatistics
Stanford University
Stanford, California
Designing a Cost-Effective Clinical Trial

John C. Hornberger, M.D., M.S.
Byron Wm. Brown, Jr., Ph.D.
Jerry Halpern, Ph.D.

Department of Health Research & Policy
Stanford University School of Medicine
Stanford, CA

January 20, 1995

PLEASE DO NOT CITE OR CIRCULATE

Correspondence to:
John C. Hornberger, M.D., M.S.
HRP Redwood Building, Room T254A
Dept of Health Research & Policy
Stanford University School of Medicine
Stanford, CA 94305-5092
(415) 725-4318
Abstract

Researchers and administrators must decide which clinical trials are worth doing and how many subjects are needed for a trial. We calculated sample size considering the costs of implementing the results of the trial and the trial costs using (1) Neyman-Pearson methods and (2) a Bayesian cost-benefit method. We illustrate these methods in a clinical trial sponsored by the National Institutes of Health that compares two levels of blood urea nitrogen clearance by hemodialysis for patients with end-stage renal disease. When applied to evaluations of research proposals, these methods may help researchers to decide whether to begin a study and, if so, how many subjects to enroll in it. These methods should be especially useful for large studies intended to inform health policy.

Keywords: Clinical trials, cost effectiveness, decision analysis, sample size, dialysis, end-stage renal disease
1 Introduction

The cost of conducting large clinical trials or epidemiologic studies intended to inform health policy can be substantial. The agencies that fund such trials therefore want to know whether benefits justify costs.

We considered the problem of estimating the required sample size for a randomized control trial under consideration by the National Institutes of Health that compares two treatment strategies for patients with end-stage renal disease.\textsuperscript{1} Controversy surrounds the decision regarding optimal administration of hemodialysis. In the past 15 years, research has shown that the removal, or clearance, of blood urea nitrogen during each dialysis session is among the best predictors of a decreased risk of morbidity.\textsuperscript{2,3} The measure of reduction used commonly in clinical practice is fractional urea clearance, or Kt/V.\textsuperscript{4} In simple kinetic models of urea nitrogen clearance from blood during dialysis, one calculates Kt/V as the -ln (BUN\textsubscript{0}/BUN\textsubscript{1}), in which BUN\textsubscript{0} and BUN\textsubscript{1} are the blood urea nitrogen levels just before starting and after a dialysis treatment, respectively. Analysis of the most comprehensive clinical trial of dialysis to date, the National Cooperative Dialysis Study, showed that levels of Kt/V less than 0.8 were associated with a 5-fold increase in hospital admission rates and uremic symptoms compared to levels of Kt/V greater than 0.8.\textsuperscript{5,6} More recent nonexperimental data collected for the U.S. Renal Data System (USRDS), a comprehensive clinical database, suggest that levels of Kt/V up to 1.4 or more may decrease patient morbidity and mortality beyond that of levels of Kt/V of 0.8-1.0.\textsuperscript{7} Patients, however, rarely receive these higher levels of Kt/V. The reluctance of some clinicians to adopt higher levels of Kt/V may reflect their uncertainty concerning the benefits associated
with these levels weighed against the additional costs for dialysis centers to acquire more equipment and apply longer dialysis treatments.8

Because the benefits of higher levels of Kt/V are unknown and because as many as 40 percent of patients have been treated with levels of Kt/V less than 0.8,9 the Renal Physicians Association (RPA) convened a working committee to develop a statement of guidelines for the appropriate prescription of hemodialysis.10 The objective of the statement was to summarize the existing literature on indications and guidelines for dialysis and to recommend strategies to improve the quality of dialysis care in the United States. The results of a cost-effectiveness analysis suggested that even small improvements in health-related outcomes (e.g., survival rate, hospital admission rates, intradialysis complications) with increasing levels of Kt/V would be cost effective.8

To provide definitive evidence for the choice of a recommended level of Kt/V in dialysis, one could consider a clinical trial; however, such a trial is costly. Is such a trial worth doing and how many subjects are needed to learn if the potential benefits from conducting a trial outweigh the trial costs? We explore this question using two approaches. We discuss the methods in Section 2 and present detailed calculations in Section 3 for the hemodialysis problem that motivated this investigation. Section 4 states the conclusions. Section 5 contains a glossary of definitions of the notation.

2 Methods

Assume that the proposed trial is a randomized, controlled, parallel-arm study with \( n/2 \) subjects in each of two treatment groups and all patients observed for 2 years or to death. The test statistic is based on the observed survival rates in the two treatment arms, \( T_1 \) and \( T_2 \), after
having observed each patient for up to 2 years. Let $P_1$ and $P_2$ denote the "true" proportions of subjects expected alive at 2 years in the control and experimental arms, respectively. We model the immediate costs of the treatments, the long-term costs of implementing each treatment, and the health-related benefits of the two treatments.

We take the survival times for the two treatments as exponentially distributed with expectations $\mu_1$ and $\mu_2$. In terms of the 2-year "true" or expected survival proportions, $P_5$, $\mu_t$ is equal to $\frac{2}{\ln P_5}$. We denote the expected cost of treating a patient with treatment $T_i$, as $C_i = C_i(P_i)$ and the quality-adjusted life expectancy for treatment $T_i$ as $Q_i = Q_i(P_i)$. We denote the maximum dollars society is willing to pay for 1 year of added quality-adjusted life by $M$. Then the overall expected cost-benefit, expressed entirely in dollars, of treating a patient with treatment $T_i$ is $S_i = S_i(P_i)$:

$$S_i(P_i) = C_i(P_i) - M \cdot Q_i(P_i).$$  \hspace{1cm} (1)

We can write the difference in the expected cost of treating a patient with the experimental treatment, $S_2(P_2)$, versus the cost for the control treatment, $S_1(P_1)$, as:

$$DS(P_1, P_2) = S_2(P_2) - S_1(P_1) = \left\{C_2(P_2) - C_1(P_1)\right\} - M \left\{Q_2(P_2) - Q_1(P_1)\right\}. \hspace{1cm} (2)$$
For $DS$ negative, $T_2$ has the advantage in expected total cost per patient and we should choose it over $T_1$. For $DS$ positive, we should choose $T_1$ over $T_2$. For $DS$ equal to 0, we are indifferent between $T_1$ and $T_2$. If the information on $P_1$ and $P_2$ is insufficiently clear to distinguish the sign of $DS$, a study might be in order, provided the cost of the study is not too great relative to the possible differences in costs over all future patients by making a wrong decision about the effectiveness of the treatments.

We assume that the study compares Kt/V at two levels, 1.2 ($T_1$, control) and 1.6 ($T_2$, experimental). Based on past experience and extensive review of the literature, we believe that the 2-year survival for the average patient in the United States treated with a Kt/V equal to 1.2 is $P_1 = 0.75$. $P_1 = 0.75$ is an approximation and the efficiency of $T_2$ as denoted by $P_2$ is far less certain. Based on the best knowledge available, we model $Q$, $S$, and $DS$ as functions of $P_1$ and $P_2$. Figure 1 shows the relationship of $DS$ to $P_1$ and $P_2$ for $M$ equal to $50,000 per added quality-adjusted life year, based on a model used to estimate cost effectiveness of Kt/V. Given $P_1$, there exists a $P_2^0 = P_2^0(P_1)$ — also a function of $M$ — where $DS(P_1, P_2^0)$ is equal to 0. This determines the minimum treatment difference, $P_2^0(P_1) - P_1$, at which we might consider shifting from the control to the experimental treatment. The more one wishes to pay for an added quality-adjusted year of life ($M$), the smaller the value of $P_2^0(P_1) - P_1$.

2.1 Neyman-Pearson I

To decide whether to adopt $T_1$ or $T_2$, we test the hypothesis $H_0$: $P_2 \leq P_2^0(P_1)$ versus $H_A$: $P_2 > P_2^0(P_1)$. $P_1$ and the cost function, $DS$, determines $P_2^0(P_1)$ and, thus, fix the study hypothesis.

We reject $H_0$ and adopt $T_2$ if:
\[ A_n = \frac{\hat{P}_2 - P_2^0 \hat{P}_1}{\hat{\delta}} > k_a, \]  

and otherwise accept \( T_1 \), where \( \hat{P}_1 \) and \( \hat{P}_2 \) are the observed proportions of 2-year survivals in the two arms, and \( \hat{\delta} \) is the estimate of the standard deviation, \( \sigma \), of the numerator of \( A_n \):

\[
\hat{\delta} = \sqrt{\frac{\hat{P}_2(1 - \hat{P}_2)}{n/2} + \left( \frac{dP_2^0}{dP_1}\right)^2 \frac{\hat{P}_1(1 - \hat{P}_1)}{n/2}}.
\]

\( A_n \) is approximately normally distributed, with a mean of zero and standard deviation of one, under \( H_0 \), and \( k_a \) is the upper 100*\( \alpha \)th percentile of the standard normal distribution.\(^*\)

To calculate an appropriate sample size \( n \) for this study, we would ordinarily specify an appropriate alternative to the null hypothesis. Here, the null hypothesis is that \( P_2 - P_2^0 (P_1) \leq 0 \) and we construct the test to yield a one-tailed type-I risk, or \( \alpha \), when \( P_2 = P_2^0 (P_1) \), i.e., when we are indifferent to the choice of \( T \) based on expected cost per patient. For \( P_1^* \), where \( P_1^* \) is our best estimate of \( P_1 \) from past experience, we select an alternative, \( P_2^* - P_2^0 (P_1^*) = \delta_n(P_1^*), > 0 \), such that \( DS(P_1, P_2^*) \) is equal to \( DS^* \), a value sufficiently negative to compel a clinician to use \( T_2 \).

\(^*\) By Taylor expansion, \( P_2^0(x) = P_2^0(\mu) + (x-\mu) \left( \frac{dP_2^0(x)}{dx} \right) \biggr|_{x=\mu} \). If \( \mu=P_1 \), then the variance of \( P_2^0(x) \) evaluated at \( P_1 \) is approximated by

\[
\sigma^2(P_2^0(\hat{P}_1)) = \left( \frac{dP_2^0}{dP_1}\right)^2 \sigma^2(\hat{P}_1),
\]

where \( \sigma^2(\hat{P}_1) = P_1(1 - P_1)/(n/2) \).
The critical effect size, which in this case we have called \( \delta_A \), typically is decided for each new study by its investigative team. Because \( \delta_A \) follows from specification of \( DS \) and \( DS^* \), \( \delta_A \) depends on the future costs of a decision based on the study results. One advantage of specifying \( DS^* \) is that it could provide a uniform criterion for planning and funding across similar studies within an organization such as the NIH.

The power, \( 1 - \beta \), at \( P_2 - P_2^0 \ (P_1) = \delta_A(P_1) \), using \( A_n \) as the test statistic, is:

\[
\text{Power} = Pr(A_n > k_r \mid P_2 = P_2^0(P_1) + \delta_A(P_1)) .
\]  

(5)

For large \( n \), the power is:

\[
\text{Power} = Pr \left( \frac{(\hat{P}_2 - P_2^0(\hat{P}_1)) - (P_2 - P_2^0(P_1))}{\sigma} > \frac{\delta_A}{\sigma} \right)
\]

\[= \Phi \left( \frac{P_2 - P_2^0(P_1)}{\sigma} - k_r \right) = \Phi(k_r) \, .
\]

(6)

where \( \Phi(k_r) \) is the upper 100*\( \beta \)th percentile of the standard normal distribution.

The sample size, \( n \), that produces the desired power at \( P_2^* = P_2^0(P_1^*) + \delta_A(P_1^*) \) is approximately:

\[
n = \frac{2(k_r + k_p)^2}{\delta_A(P_1^*)^2} \left[ P_2^*(1-P_2^*) + \left( \begin{array}{c} dP_2^0 \left( \frac{dP_2}{dP_1} \right) \end{array} \right)^2 P_1^*(1-P_1^*) \right].
\]

(7)
Note that with equation 3, we are testing if \( P_2 > P_2^0(P) + \delta_{a'}(P) \) instead of \( P_2 > P_1 + \delta_{a'} \) for some arbitrary choice of \( \delta_{a'} \) as is usual. Thus, equation 7 above explicitly takes into account costs of error in the choice of \( T_1 \) versus \( T_2 \). Because \( P^* \) is not precisely known, it is a good practice to determine the required sample size for several possible values of \( P \).

2.2 Neyman-Pearson II

In Section 2.1 above, we chose the null hypothesis at \( P_2^0 = P_2^0(\varphi) \). We chose the alternative hypothesis, \( P_2^* = P_2^0(\varphi) + \delta_{a}(\varphi) \), at the point where we deem the loss, in choosing \( T_1 \) erroneously to be too large, i.e. at \(-DS(P, P_2^*(\varphi)) = DS^* \). It might be more reasonable to choose the null hypothesis similarly, i.e., as \( P_2^{**} = P_2^0(\varphi) - \delta_{a}(\varphi) \), such that \( DS(P, P_2^{**}) \) is equal to \( DS^{**} \) (e.g., equal to \( DS^* \)), a value sufficiently positive to compel the clinician to continue to use \( T_1 \). Thus, the interval between \( P_2^0(\varphi) - \delta_{a}(\varphi) \) to \( P_2^0(\varphi) + \delta_{a}(\varphi) \) is the interval of the clinician's indifference to the choice of \( T \) for a given \( P \). The calculation of \( n \) proceeds as before, for a specific \( \alpha \) and \( \beta \), based on the best estimate of \( P \), e.g., \( P_i = P_i^* \). In this case the test statistic is:

\[
B_n = \frac{\hat{P}_2 - P_2^0(\hat{P}_1)}{\hat{\delta}} > k_n; \tag{8}
\]

and \( \hat{\delta} \), the estimate of the standard deviation of the numerator of \( B_n \) is:

\[
\hat{\delta} = \sqrt{\frac{\hat{P}_2(1 - \hat{P}_2)}{n/2} + \left( \frac{dP_2^0}{dP_1} \right)^2 \left( \frac{d\delta_a}{dP_1} \right)^2 \frac{\hat{P}_1(1 - \hat{P}_1)}{n/2}}. \tag{9}
\]
The power of a test to detect a significant enough difference to compel the use of $T_2$ at the alternative is

$$\text{Power} = \Pr(B_n > k_n \mid P_2 = P_2^0(P_1) + \delta_4(P_1))$$

$$= \Pr\left( \frac{(\hat{P}_2 - P_2^0(P_1) + \delta_4(P_1)) - (P_2 - P_2^0(P_1) + \delta_4(P_1))}{\sigma} > k_n - \frac{(P_2 - P_2^0(P_1) + \delta_4(P_1))}{\sigma} \right). \quad (10)$$

Again, assuming we use the best estimate of $P_1 = P_1^*$, with the alternative of interest equal to $P_2^* = P_2^0(P_1^*) + \delta_4(P_1^*)$, using $B_n$ as the test statistic, the sample size $n$ is approximately:

$$n = \frac{2(k_n + k_0)^2}{\left[ \delta_0(P_1^*) + \delta_4(P_1^*) \right]^2} \left[ P_2^*(1 - P_2^*) + \left( \frac{dP_2^0}{dP_1^*} \right)^2 + \left( \frac{d\delta_0}{dP_1^*} \right)^2 \right] P_1^*(1 - P_1^*). \quad (11)$$

To accommodate the uncertainty about $P_1$, one could repeat the calculation as before for several possible $P_1$'s and find the corresponding $n$ for each of them.

### 2.3 Bayesian Method

We also calculate sample size using a Bayesian cost-benefit method. For any patient, the expected loss evaluated at $P_1$ and $P_2$, $L_1(P_1, P_2)$, for using $T_1$ when we should have used $T_2$, i.e., when $P_2 > P_2^0(P_1)$, is $-DS(P_1, P_2)$. The expected loss evaluated at $P_1$ and $P_2$, $L_2(P_1, P_2)$, for using
\( T_2 \) when we should have used \( T_1 \), i.e., \( P_2 < P_2^0(P_1) \), is \( DS(P_1, P_2) \). We assume no loss for correctly using \( T_1 \) or \( T_2 \). We denote the number of persons who may benefit from the results of the trial as \( N \) and the expected cost per subject enrolled in the trial as \( C_3 \). We obtain from the trial the estimated loss of adopting \( T_n \), based on the observed \( \hat{P}_1 \) and \( \hat{P}_2 \), called the posterior expected loss for \( T_1 \):

\[
\epsilon_{\lambda}(\hat{P}_1, \hat{P}_2) = N \left[ \int_{0}^{1} \int_{0}^{1} L_{\lambda}(P_1, P_2) h(P_1, P_2 | \hat{P}_1, \hat{P}_2) dP_1 dP_2 \right] + n C_3,
\]

(12)

where \( h(P_1, P_2 | \hat{P}_1, \hat{P}_2) \) is the joint conditional density of \( P_1 \) and \( P_2 \) given \( \hat{P}_1 \) and \( \hat{P}_2 \). The first term on the right-hand side of equation 12 represents the total cost to the target population of incorrect use of \( T_n \), given the trial results \( \hat{P}_1 \) and \( \hat{P}_2 \), based on \( n \) observations. The second term on the right-hand side of equation 12 represents the cost of the trial. The Bayes Decision Rule, \( d_{n}^*(\hat{P}_1, \hat{P}_2) \), is to use \( T_1 \) if \( \epsilon_{1,n} < \epsilon_{2,n} \), use \( T_2 \) if \( \epsilon_{1,n} > \epsilon_{2,n} \), and use either \( T_1 \) or \( T_2 \) if \( \epsilon_{1,n} = \epsilon_{2,n} \). Hence, the posterior expected loss associated with choosing a treatment based on this Bayes Decision Rule, \( d_{n}^*(\hat{P}_1, \hat{P}_2) \), is equal to the \( T_i \) that corresponds to the minimum of \( \epsilon_{1,n} \) and \( \epsilon_{2,n} \) for any observed \( \hat{P}_1 \) and \( \hat{P}_2 \). For a target population with \( P_1 \) and \( P_2 \) for the two treatments, the expected loss for a study of sample size \( n \) is:

\[
R_n(P_1, P_2) = \int_{0}^{1} \int_{0}^{1} L_{d_{n}^*}(P_1, P_2) f_n(\hat{P}_1, \hat{P}_2 | P_1, P_2) d\hat{P}_1 d\hat{P}_2,
\]

(13)
where \( f_\ast(\hat{P}_1, \hat{P}_2 | P_1, P_2) \) is the joint conditional density of \( \hat{P}_1 \) and \( \hat{P}_2 \) given \( P_1 \) and \( P_2 \). Upon planning a study, the researchers express their best estimate of the distribution of \( P_1 \) and \( P_2 \), i.e. their prior distribution, as a joint density function, \( g(P_1, P_2) \). The Bayes Risk of a study with sample size \( n \), \( R_n \) is then:

\[
R_n = N \left[ \int_0^1 \int_0^1 R_\ast(P_1, P_2) g(P_1, P_2) dP_1 dP_2 \right] + n C_n \tag{14}
\]

The Bayes sample size is the sample size \( n \) that minimizes \( R_n \).

3 Illustration

A decision analysis was done for the development of a clinical guideline on the optimal prescription of hemodialysis.\(^8\) We illustrate these methods for calculating an "optimal" sample size using a cost function, \( DS \), based on that analysis.

3.1 Assumptions

For the Neyman-Pearson methods I and II, we calculated sample sizes with a Type 1 (\( \alpha \)) risk of 0.05 (one tail) and a Type 2 (\( \beta \)) risk of 0.2.

Currently, there are about 140,000 persons per year who undergo hemodialysis in the United States.\(^7\) If we assume the study will have implications for the treatment of patients who undergo hemodialysis for at least 10 years after completion of the study, then as many as 1 to 2
million persons in the United States may be affected by the study results and policy recommendations.

The number of persons who would benefit from the study depends on when new technologies replace existing ones. We assumed that 2 million persons, $N$, could benefit from the trial. We set the cost per subject in the trial equal to the budgeted average cost per subject in the proposed NIH study, i.e. $8,000 per subject. We chose $M$ equal to $50,000 per quality-adjusted life-year saved.

We assumed that the prior density function, $g$, is approximately bivariate normal \(-BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\). We let $\mu_1$ and $\mu_2$ equal 0.75, indicating that the researchers wished to design a trial sufficiently powerful to persuade skeptical clinicians who believed that $T_2$ had no advantage, on average, in survival over $T_1$. We calculated $\sigma_1^2$ such that $\int_{0.78}^{1} g(P_1) dP_1 = 0.05$ and $\sigma_2^2$ such that $\int_{0.83}^{1} g(P_2) dP_2 = 0.05$, where $g(.)$ denotes the marginal density of $P_i$. With regard to the correlation, we let $dP_2/dP_1$ represent the researchers' belief about the expected change in $P_2$ for a unit change in $P_1$. Then, using the conditional normal distribution of $P_2$ given $P_1$, $\rho = (\sigma_1/\sigma_2) (dP_2/dP_1)$. We assumed for this exercise that $dP_2/dP_1$ equaled 0.5, i.e., that we expect a 0.10 increase in $P_1$ corresponds with an increase of 0.05 in $P_2$. Taking $\hat{P}_1$ and $\hat{P}_2$ as independent, conditional on $P_1$ and $P_2$, the mean of $f_n(\hat{P}_i|P_1)$ is $P_i$, with variance $P_i(1 - P_i)/(n/2)$.

\[\text{\dag} \text{ Using a bivariate normal distribution, } P_i | P_j - N(\mu_i + \rho (\sigma_i/\sigma_j) (P_j - \mu_j), \sigma_i^2 (1 - \rho^2)); \text{ hence, } P_2 = \mu_2 + \rho (\sigma_2/\sigma_j) (P_1 - \mu_j) \text{ and } dP_2/dP_1 = \rho (\sigma_2/\sigma_1). \text{ We then calculated } \rho \text{ as } (\sigma_1/\sigma_2) (dP_2/dP_1).\]
To obtain the Bayes Decision Rule for selecting $T_j$, where:

$$
\epsilon_{1n} - \epsilon_{2n} = N \left[ \int_{0}^{1} \int_{0}^{1} DS(P_1, P_2) g(P_1, P_2) f_n(P_1, P_2 | P_1, P_2) dP_1 dP_2 \right],
$$

and using numerical integration, we found the level of $\hat{P}_2, \hat{P}_{2^*}$, such that $\epsilon_{1n} - \epsilon_{2n} = 0$ for $\hat{P}_1 = 0.71$ (0.01) 0.80, for $n = 1000$ (200) 8000. The Bayes Risk for a given target population, with $P_1$, $P_2$, and $n$ subjects then is

$$
R_n(P_1, P_2) = \begin{cases} 
-DS(P_1, P_2) Pr(\hat{P}_2 < \hat{P}_{2^*}(\hat{P}_1, n)) & \text{if } P_2 \in (P_2^0, 1) \\
DS(P_1, P_2) (1 - Pr(\hat{P}_2 < \hat{P}_{2^*}(\hat{P}_1, n))) & \text{if } P_2 \in (0, P_2^0) 
\end{cases}
$$

and the Bayes Risk of a study with $n$ subjects is

$$
R_n = N \left[ \int_{0}^{1} \int_{P_2^0}^{1} -DS(P_1, P_2) g(P_1, P_2) Pr(\hat{P}_2 < \hat{P}_{2^*}(\hat{P}_1, n)) dP_2 dP_1 
\right. \\
\left. + \int_{0}^{1} \int_{P_2^0}^{1} DS(P_1, P_2) g(P_1, P_2) (1 - Pr(\hat{P}_2 < \hat{P}_{2^*}(\hat{P}_1, n))) dP_2 dP_1 \right] + nC_s
$$

We found the $n$ that minimized $R_n$ by a search over $n$. 

13
3.2 Results

Table 1a shows the sample sizes calculated by the Neyman-Pearson Method I for several values of $DS^*$. Table 1b shows the sample sizes calculated by the Neyman-Pearson Method II for several values of $DS^{**} = DS^*$ where $P_1 = 0.75$ and the point of indifference between $T_1$ and $T_2$ is $P_2 = 0.83$.

The larger the value of $DS^*$ (i.e., the greater the risk we are willing to accept that the trial will incorrectly reject $T_2$), the larger the critical effect sizes, $\delta_0$ and $\delta_A$, and the smaller the sample size that is needed for the trial. For Method I, if the $DS^* = $5,000, then $\delta_A$ is 0.012 and the optimal sample size is 16,046. If, instead, $DS^* = $15,000 then $\delta_A$ is 0.032 and the optimal sample size is only 2,641. For a given $DS^{**} = DS^* = $5,000, $\delta_A$ is the same for Methods I and II, but the size of the region of indifference increases from $\delta_A - P_2^0 = 0.012$ for Method I to $\delta_A - \delta_0 = 0.035$ for Method II. Thus, for a given $DS^*$, the sample size for Method II is smaller than that for Method I because we are trying to discriminate between the upper and lower ends of the indifference region, and the indifference region is longer for Method II than for Method I.

The expected costs appear in the last columns of Tables 1a and 1b. These costs are obtained by replacing $d_n^*$ in equation 13 by the decisions about which treatment to use based on estimates of $A_n$ in equation 3 and $B_n$ in equation 8, respectively. In Method II, placing the point of indifference roughly halfway between the hypothesized $P_2^{**}$ and the alternative $P_2^*$ and using the decision rule in equation 8 with $\alpha = 0.05$ and $\beta = 0.2$, the sample size that minimizes the expected cost is about $32 million in trial costs (8,000 per subject * 4,000 subjects). Other $DS^*$ result in much larger expected study costs.

At the estimated survival rate of 75 percent for the control group, the Bayes Rule yields an optimal sample size of 6,200 subjects and a trial cost of $52 million. The total expected cost
is $111 million. Thus, the Bayes method gives a design with a substantially lower cost than either of the Neyman-Pearson approaches.

4 Discussion

The number of research proposals for clinical trials submitted to federal agencies continues to rise faster than the money available to fund them. Federal advisory councils and other administrators of research monies therefore confront difficult choices about what projects to fund. Review committees also must comment on whether the design of a promising project is efficient.

Because these decisions weigh the potential gains in clinical benefits against the costs of the project and future implementation, we took the view, similar to that of Detsky and others, that explicit calculation of the expected costs and benefits of the trial can help in making the decision. The sample size estimation method presented here permits calculation of the Bayes Risk based on a nonlinear loss function.

The results of the analysis depend on the assumed prior distribution and the loss function used. The researcher needs to justify any proposed prior distribution and loss functions to the review committee. For example, the researcher needs to review the literature extensively, perhaps summarizing results using standardized techniques, such as meta- or decision-analyses. One might survey experts or practicing clinicians to provide an estimate of the prior probability distributions. The researcher could alter the parameters of the prior distribution to learn how results vary with such changes. The review committee itself may assign a different prior probability to learn how their beliefs affect the calculated sample size. The Bayes approach
reduces the guesswork and, at best, makes the role of such guesses more explicit to the researchers and to reviewers of the trial.

While these efforts will add costs to the design of the study, they may be justified for particularly expensive proposals. For example, the cost of the Diabetes Control and Complications Trial (DCCT)\textsuperscript{20} in patients with type-I diabetes cost in excess of $150 million. It is likely that a similar investment might be necessary to learn whether the same interventions used in the DCCT are warranted for patients with type-II diabetes. An investment of less than one percent of the trial budget to perform the steps proposed here may be reasonable to avoid collection of more data than needed to establish the effectiveness of the intervention and to assure that there are sufficient data collected to avoid missing an important effect.

By conducting a cost-benefit analysis before funding of a trial, we anticipate that one could use the analysis to formulate a written guideline — as was done for the RPA guideline on the adequacy of hemodialysis — to inform physicians with the most up-to-date information about the intervention. Many large trials sponsored by the federal government evaluate existing interventions already in clinical use. We illustrated these methods with a proposed study of dialysis technologies currently widely available to clinicians. Other examples abound in medicine, including studies of tPA versus streptokinase in acute myocardial infarction,\textsuperscript{21} aspirin versus warfarin for non-rheumatic atrial fibrillation,\textsuperscript{22} and dietary modification to prevent progression of renal disease.\textsuperscript{23} As findings of such trials reach publication, one could use the cost-benefit models developed for designing the study to update estimates of the benefits and costs of these interventions and thereby to improve on the guidelines.

We found that sponsorship of the proposed dialysis study is a reasonable decision. It is possible for some problems that there exists no sample size that justifies the costs of the trial —
i.e., the optimal $n = 0$. The analysis therefore not only informs the researcher about the desired
sample size, but also indicates whether the study is worth doing at all.

Our approach requires specification of value for variables for which there may be little
information, such as the size of the target population. This value depends inherently on the rate
that competing strategies will be developed and the expected incidence of the disease over at
least the next 5 to 10 years. We found little data to guide our estimate of the potential size of the
target population. It is wrong, however, to conclude that because such values are uncertain that
the method is of little use. Many clinical decisions require that the analyst include variables for
which there is little information. The point is to bring the available information, such as it is, to
bear on the problem at hand in as effective a way as possible. As in specifying a prior
distribution on the effect size, the research team may survey experts about the likely size of the
target population or perform sensitivity analyses to assess the effect of changes in this variable
on the estimated sample size.
5 Glossary

Kt/V Fractional urea clearance.
BUN Blood urea nitrogen.
$T_i$ Treatment options ($i = 1$ or $2$).
$P_i$ 2nd-year survival for the $i$th treatment.
$\mu_i$ Expectations of survival time for the $i$th treatment.
$\hat{P}_i$ Observed 2-year survival for $i$th treatment.
$P_i^0(P_j)$ Minimum cost-effective effect size.
$Q_i$ Quality-adjusted life expectancy for $i$th treatment.
$C_i$ Lifetime medical costs for $i$th treatment.
$S_i$ Overall cost benefit in dollars for $i$th treatment.
$DS(P_1, P_2)$ Difference in overall cost benefit between the control and experimental treatments.

$A_n$ Test statistic for Neyman-Pearson I method.
$B_n$ Test statistic for Neyman-Pearson II method.
n Desired sample size.
$DS^*$ Level of $-DS$ to compel use of $T_1$ with Neyman-Pearson I method.
$DS^{**}$ Level of $DS$ to compel use of $T_1$ with Neyman-Pearson II method.
$C_N$ Expected cost per subject in the trial.
$N$ Expected size of the target population.
$M$ Maximum dollars society is willing to pay to prolong life by 1 quality-adjusted life year.

$\alpha$ Type-I error.
$\beta$ Type-II error.
$k_x$ $(1-x)$ 100th percentile of standard normal distribution.
$\Phi$ Cumulative standard distribution function.
$\delta_i(P_j)$ The difference in $P_2$ and $P_2^0(P_j)$ defining the alternative of interest in the Neyman-Pearson methods, i.e., where $DS = DS^*$ and $P_2 > P_2^0(P_j)$.
$\delta_0(P_j)$ The difference in $P_2$ and $P_2^0(P_j)$ defining the null hypothesis in the Neyman-Pearson method II, i.e., where $DS = DS^*$ and $P_2 < P_2^0(P_j)$.
$g(P_1, P_2)$ Joint density of $P_1$ and $P_2$.
$h(P_1, P_2 | \hat{P}_1, \hat{P}_2)$ Conditional density of $P_1$ and $P_2$ given $\hat{P}_1$ and $\hat{P}_2$.
$f(\hat{P}_1, \hat{P}_2 | P_1, P_2)$ Conditional density of $\hat{P}_1$ and $\hat{P}_2$ given $P_1$ and $P_2$.
$L_i(P_1, P_2)$ Bayesian loss function of the $i$th treatment evaluated at $P_1$ and $P_2$.
$\epsilon_{w}(\hat{P}_1, \hat{P}_2)$ Posterior expected loss of the $i$th treatment for a trial of sample size $n$ evaluated at $\hat{P}_1$ and $\hat{P}_2$.

$D_2^*(\hat{P}_1, \hat{P}_2)$ Bayesian Decision Risk for given $\hat{P}_1$ and $\hat{P}_2$.
$\hat{P}_2(P_1, n)$ The $\hat{P}_2$ such that $\epsilon_2 - \epsilon_1 = 0$ at $\hat{P}_1$ and $n$.
$R_n(P_1, P_2)$ Bayes risk for trial of sample size $n$ evaluated at $P_1$ and $P_2$.
$R_n$ Bayes risk for a trial with sample size $n$. 

18
6 Acknowledgements

Dr. Hornberger is a Picker/Commonwealth Scholar. This study was supported, in part, by an NIH grants, GMZ1215-17, R01 DK47805-01, 9 R01 CA59039-18, and The Center for Health Care Evaluation, Department of Veterans Affairs, VA Medical Center, Palo Alto, CA. Some of the data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as the official policy or interpretation of the Canadian government or the U.S. government. The authors wish to thank Melissa Smith for report preparation.
7 References


Table 1a. Study designs for Neyman-Pearson Method I

<table>
<thead>
<tr>
<th>DS*(D)</th>
<th>P₂*(P₁*)+δₐ(P₁*)</th>
<th>n</th>
<th>Expected Cost ($, million)³</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>0.842</td>
<td>16,046</td>
<td>204</td>
</tr>
<tr>
<td>10,000</td>
<td>0.853</td>
<td>4,978</td>
<td>226</td>
</tr>
<tr>
<td>15,000</td>
<td>0.862</td>
<td>2,641</td>
<td>304</td>
</tr>
<tr>
<td>20,000</td>
<td>0.868</td>
<td>1,880</td>
<td>368</td>
</tr>
</tbody>
</table>

³ P₁*=0.75, P₂*(P₁*)=0.83, α=0.05, β=0.2.

b Expected costs replacing dₙ* with estimates of Aₙ (equation 3).

Table 1b. Study designs for Neyman-Pearson Method II

<table>
<thead>
<tr>
<th>DS**(D)</th>
<th>P₂*(P₁*)-δₐ(P₁*)</th>
<th>P₂*(P₁*)+δₐ(P₁*)</th>
<th>n</th>
<th>Expected Cost ($, million)³</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,000</td>
<td>0.816</td>
<td>0.837</td>
<td>10,900</td>
<td>272</td>
</tr>
<tr>
<td>4,000</td>
<td>0.812</td>
<td>0.840</td>
<td>6,100</td>
<td>238</td>
</tr>
<tr>
<td>5,000</td>
<td>0.807</td>
<td>0.842</td>
<td>3,890</td>
<td>237</td>
</tr>
<tr>
<td>6,000</td>
<td>0.802</td>
<td>0.844</td>
<td>2,466</td>
<td>251</td>
</tr>
<tr>
<td>7,000</td>
<td>0.796</td>
<td>0.847</td>
<td>1,682</td>
<td>279</td>
</tr>
<tr>
<td>8,000</td>
<td>0.790</td>
<td>0.849</td>
<td>1,224</td>
<td>322</td>
</tr>
</tbody>
</table>

³ P₁*=0.75, P₂*(P₁*)=0.83, α=0.05, β=0.2.

b Expected costs replacing dₙ* with estimates of Bₙ (equation 8).
Figure 1. $DS$ of $P_1$ and $P_2$. Solid line in $DS$ approximates the points where $DS$ equals 0.