Using Prior Opinions to Examine Sample Size in Two Clinical Trials

Chin-pei Tsai and Kathryn Chaloner

School of Statistics, University of Minnesota
Technical Report No. 635

October, 1999

Using Prior Opinions to Examine Sample Size in Two Clinical Trials

Chin-pei Tsai
Kathryn Chaloner

ABSTRACT Two examples of large clinical trials for the treatment of advanced HIV disease are described. Chaloner and Rhame (1999) elicited prior opinions about the outcomes of the two trials from over 50 HIV clinicians. Their prior opinions are used here for design: the sample size for reaching consensus with high probability is calculated. Consensus is said to occur when all clinicians have posterior opinions which would lead to prescribing the same treatment. Posterior beliefs are calculated using a simple linear Bayes approximation. In addition plots are given for determining parameter values for which a particular sample size is sufficient for consensus to be reached with high probability. These calculations are useful tools at the design stage and are simple to implement.

1 Introduction

The design of two large AIDS trials is examined. The trials were designed by the Community Program for Clinical Research in AIDS (CPCRA) a collaborative group, sponsored by the U.S. National Institutes of Health. Both trials examined long term treatment for the prevention of a common opportunistic infection, *Pneumocystis carinii* pneumonia (PCP) in patients with advanced HIV disease. For the two trials, Chaloner and Rhame elicited prior opinions from over 50 clinicians. These prior opinions are used here and details of the elicitation, with other uses of the opinions, are in Chaloner and Rhame (1999).

Both these trials were large pragmatic trials aimed to influence clinical practice. The approach we take, therefore, is that when the trial is done, the data should be sufficient so that clinicians will generally have consensus about what treatment to prescribe. Having the opinions elicited by Chaloner and Rhame for these two trials enables these calculations to be done for these two trials.

2 Assumptions and Notation

Suppose the two trials are both balanced designs with n independent Bernoulli observations on each treatment denoted $\{X_n\}$ and $\{Y_n\}$. The probabilities

of PCP in two years are θ_X and θ_Y, respectively. $\bar{X}$ and $\bar{Y}$ are the sample proportions for the two treatments. An normal approximation will be used to the posterior distribution of θ_X and θ_Y given the data, $\bar{X}$ and $\bar{Y}$. A different approximation, using the results of Hartigan (1969), was used in Tsai (1999). Hartigan's approximation requires prior means and variances for θ_X and θ_Y but Chaloner and Rhame elicited prior means for θ_X and θ_Y and only a prior variance for $\theta_X - \theta_Y$. The approximation used here is a normal approximation to the marginal posterior distribution for $\theta_X - \theta_Y$ given the data $\bar{X}$ and $\bar{Y}$.

Specifically, denote the prior mean of $\theta_X - \theta_Y$ as $m_X - m_Y$ and the prior variance as τ^2. The distribution of $\bar{X} - \bar{Y}$, given θ_X and θ_Y, is approximately normal with mean $\theta_X - \theta_Y$ and variance $n^{-1}\sigma^2$ where $\sigma^2 = \theta_X(1-\theta_X) + \theta_Y(1-\theta_Y)$. So approximating both the prior distribution and the likelihood with that of a normal distribution gives a posterior mean, for $\theta_X - \theta_Y$, denoted $\mu(\bar{X}, \bar{Y})$, of

$$\begin{aligned} & \mu(\bar{X}, \bar{Y}) \\ \approx & \frac{n\tau^2}{n\tau^2 + \sigma^2}(\bar{X} - \bar{Y}) + \frac{\sigma^2}{n\tau^2 + \sigma^2}(m_X - m_Y). \end{aligned}$$

Given θ_X and θ_Y, therefore,

$$\begin{aligned} Pr[\mu(\bar{X}, \bar{Y}) \leq k | \theta_X, \theta_Y] &\approx Pr\left[\bar{X} - \bar{Y} \leq \frac{k(n\tau^2 + \sigma^2) - \sigma^2(m_X - m_Y)}{n\tau^2} | \theta_X, \theta_Y\right] \\ &\approx \Phi\left(\frac{K - (\theta_X - \theta_Y)}{\sqrt{n^{-1}\sigma^2}}\right), \end{aligned}$$

where

$$K = \frac{k(n\tau^2 + \sigma^2) - \sigma^2(m_X - m_Y)}{n\tau^2}.$$

This is an approximation to the sampling probability, given θ_X and θ_Y, that the posterior mean will be less than or equal to k. For the two trials Chaloner and Rhame elicited only the means for θ_X and θ_Y and a 95% interval for $\theta_X - \theta_Y$. It is therefore assumed that if L is the length of the 95% interval then $L = 4\tau$, or $\tau^2 = L^2/16$.

2.1 Consensus

Suppose opinions are documented from t clinicians. Consensus for prescribing X is defined, in the two examples, as all t clinicians having posterior means for $\theta_X - \theta_Y$, $\mu_i(\bar{X}, \bar{Y}), i = 1 \ldots t$, less than or equal to some value k.

Similarly they will agree in prescribing Y if $\mu_i(\bar{X}, \bar{Y}) > k, i = 1 \ldots t$. The value of k must be determined from the clinical context: $k = 0$ if treatments X and Y are similar in cost, toxicity and ease of adherence.

For fixed θ_X and θ_Y, where $\theta_X - \theta_Y \leq k$, denote by n_X the sample size needed to convince all clinicians to prescribe X. Similarly for $\theta_X - \theta_Y > k$ denote n_Y to be the sample size needed to convince all clinicians to prescribe Y. Specifically define n_X to be the smallest sample size such that the probability, under the joint sampling distribution of $\bar{X}$ and $\bar{Y}$, is at least $1 - \beta$ that $\mu_i(\bar{X}, \bar{Y}) \leq k$ for all $i = 1, \ldots, t$. Similarly n_Y is the smallest sample size such that the probability, is at least $1 - \beta$ that $\mu_i(\bar{X}, \bar{Y}) > k$ for all $i = 1, \ldots, t$.

Under the sampling distribution of $\bar{X}, \bar{Y}$ the events $\mu_i(\bar{X}, \bar{Y}) \leq k$ are dependent and n_X will be the maximum of the individual sample sizes, n_i. n_i is the smallest n such that the probability that $\mu_i(\bar{X}, \bar{Y}) \leq k$ is at least $1 - \beta$.

Note that the sampling distribution of $\bar{X}, \bar{Y}$ for fixed θ_X, θ_Y is used, not the predictive distribution of $\bar{X}, \bar{Y}$. Also the value σ^2 is specified from the values of θ_X, θ_Y. Both n_X and n_Y therefore depend on θ_X and θ_Y.

3 The PCP Prophylaxis trials

The CPCRA PCP-TMS trial compares two dosing regimens of *trimethoprim sulfamethoxole* (TMP-SMX) in HIV infected patients who are not known to be intolerant of the drug. TMP-SMX is believed to be the most effective drug that can prevent PCP, the most common infection in patients with advanced HIV disease. The standard dose is one double strength tablet daily (D) but a lower dose of a double strength tablet three times a week (T) was considered. The three times a week dosing could be associated with fewer side effects and toxicities and so may therefore be tolerable for longer and so more effective. Alternatively, it could be less effective as first, bioavailability will be less, and second, patients might find it harder to remember to take a dose just three times a week.

A second trial was designed for patients who develop intolerance to TMP-SMX. This PCP-INT2 trial compares two drugs, atovaquone and dapsone, in patients who are intolerant of TMP-SMX. Dapsone was the standard treatment for patients intolerant of TMP-SMX and atovaquone was a newly licensed drug, approved, and sometimes used, for treatment of PCP, but rarely used for preventive treatment. Insurance carriers typically would not cover atovaquone for the prevention of PCP and so clinicians would not prescribe it.

For each trial, 58 clinicians were asked for their opinion about the proportion of PCP after two years for each treatment, and a 95% interval for the difference of proportions (Chaloner and Rhame, 1998). There are more

details of the elicitation in their paper.

Figure 1 is a plot of the elicited 95% intervals. Note that a complete joint probability distribution for the two probabilities was not specified.

4 The PCP-TMS Trial

In this trial, daily (D) and three times a week (T) of TMP-SMX are compared. Replacing X by D and Y by T in the previous notation let θ_D and θ_T be the probabilities of PCP in the two years after randomization. The discussion in the protocol argues that a difference of $\theta_D - \theta_T$ between $[-0.009, 0.009]$ is clinically unimportant and the two treatments would be deemed equivalent. As daily is easier for patients to remember, and the drug is extremely inexpensive daily dosing would be prescribed in this case. A clinician would therefore prescribe daily dosing if he or she had a posterior mean for θ less than or equal to 0.009; otherwise, one double strength tablet three times a week would be prescribed. So the k of Section 2.1 is 0.009.

The upper plot in Figure 1 shows the prior opinions on θ for this trial. Among the 58 clinicians, 5 of them did not provide intervals and 53 prior beliefs are therefore used to calculate the sample size. There are only two clinicians who have a prior opinion that T is clinically better than D: that is $m_D - m_T > 0.009$.

4.1 Results for the PCP-TMS trial

For the 53 prior distributions, n_D is calculated by convincing those clinicians who have $m_D - m_T$ greater than 0.009. That is, under the joint sampling distribution of $\bar{D}$ and $\bar{T}$, given θ_D and θ_T, the optimal sample size n_D is the smallest n such that the probability that all the posterior means of $\theta_D - \theta_T \leq 0.009$ is at least $1 - \beta$.

n_D and n_T are calculated numerically. Tables 1 and 2 give the sample sizes for $\beta = 0.2$ and a range of values of θ_D and θ_T. Note that the sample size is very big when $\theta_D - \theta_T$ is only a little smaller than 0.009. For example, a sample size of 89,031 for each treatment is needed to convince all 53 clinicians when θ_D is 0.212 and θ_T is 0.205. If $\theta_D = 0.092$ and $\theta_T = 0.123$, daily dosing would be considered to be meaningfully better and a sample size of 741 for each treatment will be sufficient to guarantee with probability at least 0.80 that all posterior means will be no greater than 0.009.

The actual sample size in this trial was chosen to be 1250 in each group, using frequentist hypothesis testing calculations. For this sample size, Figure 2 gives the values of θ_D and θ_T for which consensus will be reached on prescribing with probability at least 0.80. The shaded region is the region in which the actual sample size of 1250 is too small to reach consensus with probability at least 0.80.

5 The PCP-INT2 trial

The PCP-INT2 trial was designed to compare two alternative drugs, dapsone (100mg PO daily) and atovaquone (1500mg PO daily). The index X is replaced by D (dapsone) and Y is replaced by A (atovaquone) in this example. In this trial, the protocol specifies that dapsone is said to be clinically better than atovaquone if $\theta_D - \theta_A$ is less than -0.06 and the range of equivalence is [-0.06,0.06]. Since dapsone is much less expensive than atovaquone, dapsone would generally be prescribed if the effects of these two drugs were found equivalent. Therefore, the optimal sample size for prescribing dapsone, denoted by n_D, would be the smallest n such that with probability at least $1 - \beta$, all posterior means of $\theta_D - \theta_A$ will be less than or equal to 0.06. Similarly, n_A is the smallest n such that all the posterior means of $\theta_D - \theta_A$ are greater than 0.06 with probability at least $1 - \beta$. In this trial, 7 clinicians did not provide information about the results of this trial. Only 51 prior beliefs are used to calculate the sample size.

5.1 Results for the PCP-INT2 trial

In the PCP-INT2 trial, there was only one clinician who believed prior to the trial that atovaquone was clinically better: $m_D - m_A$ was bigger than 0.06. The sample size is therefore smaller for reaching a consensus for prescribing dapsone than the sample size needed for reaching consensus for prescribing atovaquone. For $\beta = 0.2$, Tables 3 and 4 show numerical results for prescribing dapsone and atovaquone for some possible values of θ_D and θ_A.

For prescribing dapsone, the sample size n_D is decreasing with θ_A as θ_D fixed and is increasing with θ_D when θ_A is fixed. If $\theta_D - \theta_A$ is less than but close to 0.06, the needed sample size is much bigger. For $\theta_D = 0.24$ and $\theta_A = 0.36$, dapsone would be considered to be meaningfully better and a sample size of 60 for each treatment is sufficient to convince the 51 clinicians.

For prescribing atovaquone, n_A is decreasing with θ_D for fixed θ_A and is increasing with θ_A for fixed θ_D. If $\theta_D = 0.24$ and $\theta_A = 0.12$ then atovaquone is considered to be meaningfully better and the sample size needed would be 12,357 for each treatment.

In this trial, the planned sample size was 700 (350 in each group). For this sample size, $n = 350$, Figure 3 shows values of θ_D and θ_A for which consensus will be reached on prescribing with probability at least 0.80. In this plot the probability of consensus is less than 0.80 when θ_D and θ_A are in the large shaded area. Note that with a sample size of 350 for each group, there is only a very small region (the dotted region) where θ_A is close to 0 and θ_D is close to 1 where consensus will be reached, with probability at least 0.80, to prescribe atovaquone. This is useful to know, and calls into question the usefulness of this trial.

6 Discussion

The sample sizes needed for prescribing the "non-standard" treatments are big for both trials. For prescribing the treatment atovaquone, in the PCP-INT2 trial, the very big sample sizes are caused by a few prior opinions which could be argued to be unreasonable. For example, one clinician has a 95 % interval of [-0.01,0.01] and this length is too small to be reasonable given that little was known about the effectiveness of the two drugs at the time. Two clinicians have means for the difference of -0.13 and -0.1 with correspondent confidence interval [-0.13,-0.07] and [-0.1,-0.01]. These two prior opinions also do not look reasonable as, again, very little data was available on the two drugs at the time the beliefs were specified but these opinions correspond to very strong opinions. It therefore seems reasonable not to consider these prior beliefs. If these three outlying beliefs are omitted, the sample sizes become much smaller.

7 Conclusions

Spiegelhalter, Freedman and Parmar (1994) and others argue that a clinical trial should stop when consensus would be reached in the scientific community about the primary result of a trial. The two trials considered here were trials with an objective of answering two questions: first what dose of TMP-SMX should be prescribed for PCP prophylaxis for patients who can tolerate TMP-SMX and second should dapsone or atovaquone be prescribed for patients who are intolerant of TMP-SMX. The methods described here are simple tools which add to considerations of sample size. They indicate that for the first question, to be answered by the PCP-TMS trial, unless the two dosages are very similar in effect, the sample size of 1250 on each treatment is probably large enough to reach consensus. For the second question if the answer is that dapsone should be used then consensus will probably be reached, but the sample size is too small to reach consensus if the answer is atovaquone. This is an important consideration for design.

Extensions to this work are in Tsai (1999). For example, an alternative approach is given there where just three prior distributions are specified, representing optimistic, pessimistic and skeptical opinions. The sample size is calculated for reaching consensus for these three prior opinions. For the two trials discussed here, however, given the wide diversity of opinion, using all the prior opinions elicited seems more appropriate.

Acknowledgments

This research was supported in part by grants from the National Security Agency and National Institutes of Health.

References

Chaloner, K. and Rhame, F. S. (1999), "Ethical and Statistical Reasons for Quantifying and Documenting Prior Opinions in Clinical Trials," Manuscript.

Hartigan, J. A. (1969), "Linear Bayesian Methods," *Journal of the Royal Statistical Soc. Ser. B* **31**, 446-454.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1994), "Bayesian Approaches to Randomized Trials (with discussion)," *Journal of the Royal Statistical Soc. Ser. A* **157**, 357-416.

Tsai, C. (1999), *Bayesian Experimental Design with Multiple Prior Distributions*, PhD Thesis, School of Statistics, University of Minnesota.

Table 1: The Optimal Sample Sizes for Prescribing Daily Dosing in the PCP-TMS Trial for Possible Values of θ_D and θ_T

		θ_T							
		0.041	0.082	0.123	0.164	0.205	0.305	0.405	0.505
	0.012	211	152	130	117	107	89	74	61
	0.032	737	260	185	153	133	104	84	68
	0.052	0	497	270	201	165	121	95	75
	0.072	0	1388	418	269	207	140	107	84
θ_D	0.092	0	0	741	374	263	163	120	92
	0.112	0	0	1896	556	342	190	135	102
	0.212	0	0	0	0	89031	494	249	165
	0.312	0	0	0	0	0	115090	594	284
	0.412	0	0	0	0	0	0	130359	641
	0.512	0	0	0	0	0	0	0	134837

Table 2: The Optimal Sample Sizes for Prescribing One Double Strength Tablet Three Times a Week in the PCP-TMS Trial for Possible Values of θ_D and θ_T

		θ_T							
		0.041	0.082	0.123	0.164	0.205	0.305	0.405	0.505
	0.012	0	0	0	0	0	0	0	0
	0.032	0	0	0	0	0	0	0	0
	0.052	65810	0	0	0	0	0	0	0
	0.072	4503	0	0	0	0	0	0	0
θ_D	0.092	2580	306329	0	0	0	0	0	0
	0.112	1923	7804	0	0	0	0	0	0
	0.212	1045	1662	2910	6918	0	0	0	0
	0.312	782	1064	1464	2088	3229	0	0	0
	0.412	622	794	1007	1285	1666	3884	0	0
	0.512	498	616	752	915	1114	1899	4197	0

Table 3: The Optimal Sample Sizes for Prescribing Dapsone in the PCP-INT2 Trial for Possible Values of θ_D and θ_A

	θ_A											
θ_D	0.04	0.09	0.14	0.19	0.24	0.29	0.34	0.39	0.44	0.49	0.54	0.59
0.04	48	33	27	24	21	19	17	15	14	13	11	10
0.09	1207	101	55	40	32	27	24	21	18	16	14	13
0.14	0	2029	148	75	51	40	32	27	24	21	18	16
0.19	0	0	2751	190	92	61	46	37	31	26	22	19
0.24	0	0	0	3373	225	106	69	51	40	33	28	23
0.29	0	0	0	0	3894	253	117	75	55	43	35	29
0.34	0	0	0	0	0	4316	276	126	80	57	44	36
0.39	0	0	0	0	0	0	4636	293	132	83	59	45
0.44	0	0	0	0	0	0	0	4857	303	135	84	59
0.49	0	0	0	0	0	0	0	0	4977	308	136	83
0.54	0	0	0	0	0	0	0	0	0	4997	306	134
0.59	0	0	0	0	0	0	0	0	0	0	4917	298

Table 4: The Optimal Sample Sizes for Prescribing Atovaquone in the PCP-INT2 Trial for Possible Values of θ_D and θ_A

	θ_A										
θ_D	0.04	0.09	0.14	0.19	0.24	0.29	0.34	0.39	0.44	0.49	0.54
0.04	0	0	0	0	0	0	0	0	0	0	0
0.09	0	0	0	0	0	0	0	0	0	0	0
0.14	10383	0	0	0	0	0	0	0	0	0	0
0.19	5431	15417	0	0	0	0	0	0	0	0	0
0.24	3963	7464	19798	0	0	0	0	0	0	0	0
0.29	3210	5166	9215	23524	0	0	0	0	0	0	0
0.34	2722	4025	6189	10683	26597	0	0	0	0	0	0
0.39	2361	3313	4708	7033	11869	29016	0	0	0	0	0
0.44	2071	2806	3799	5260	7697	12773	30781	0	0	0	0
0.49	1824	2413	3164	4183	5681	8181	13394	31893	0	0	0
0.54	1606	2090	2682	3438	4462	5970	8486	13733	32351	0	0
0.59	1408	1813	2292	2878	3626	4638	6127	8612	13789	32154	0

Figure 1: 95% prior belief intervals for both PCP-TMS trial and PCP-INT2 trial. Dotted vertical line shows the value of k.

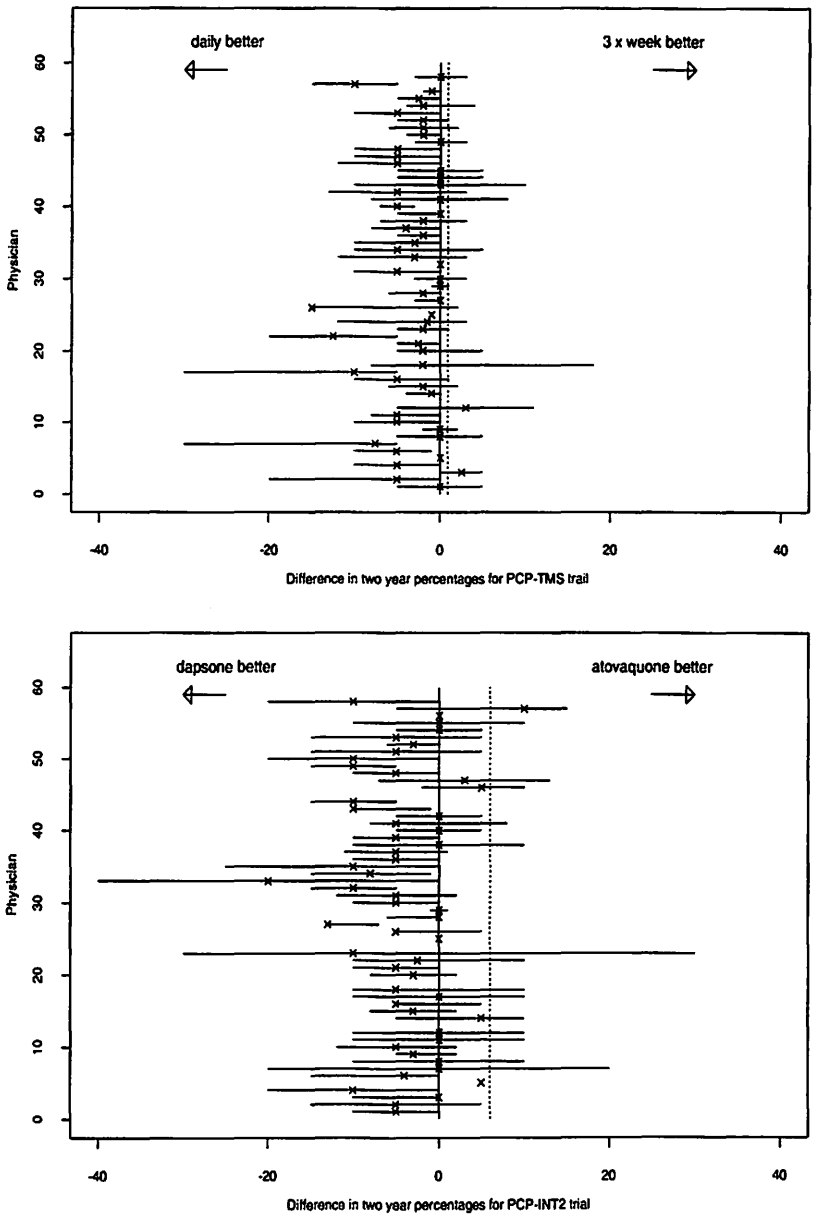

Figure 2: Values of θ_D and θ_T for which the probability of consensus on prescribing is at least 0.80 when n is 1250 for each group in PCP-TMS trial. For n=1250, the probability of consensus is less than 0.80 if θ_D and θ_T lie in the shaded area.

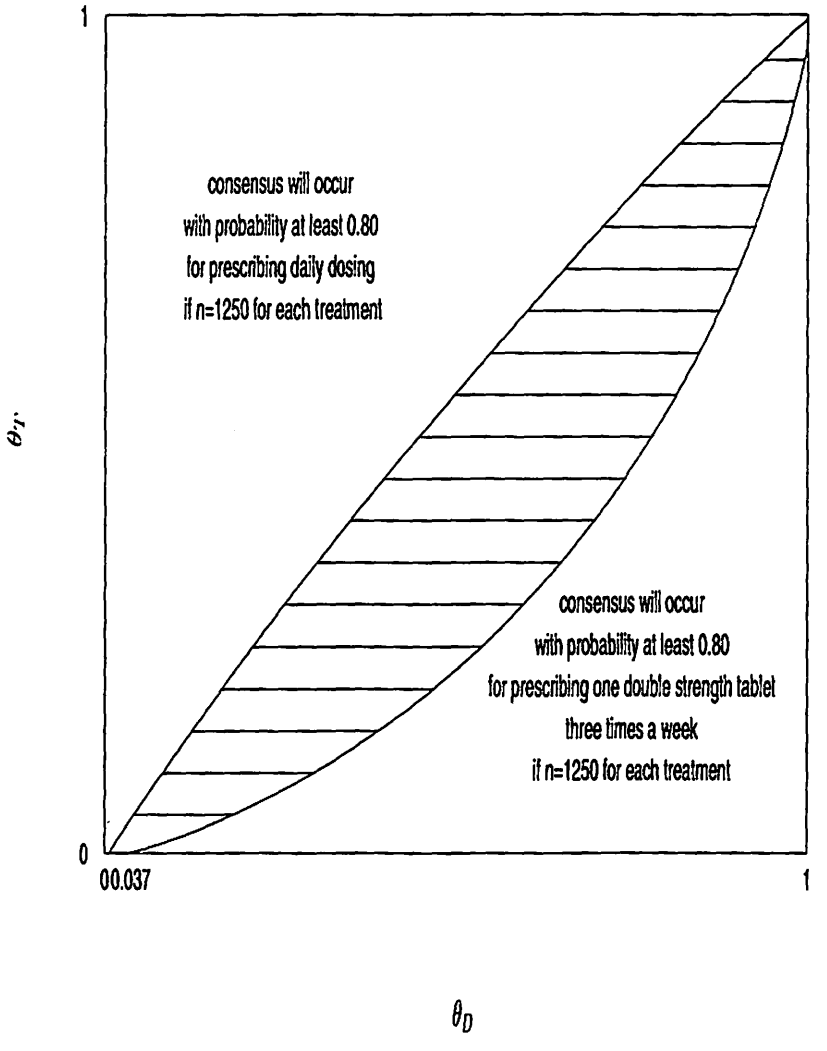

Figure 3: Values of θ_D and θ_A for which the probability of consensus on prescribing is at least 0.80 when n is 350 for each group in PCP-INT2 trial. For n=350, the probability of consensus on prescribing Atovaquone is at least 0.80 if θ_D and θ_T lie in the dotted area. If θ_D and θ_A lie in the shaded area, the actual sample size of 350 is too small to reach consensus with probability at least 0.80.