

Evaluating traditional Chinese medicine using modern clinical trial design and statistical methodology: Application to a randomized controlled acupuncture trial

Lixing Lao^{a,*†}, Yi Huang^b, Chiguang Feng^{b,c}, Brian M. Berman^a
and Ming T. Tan^{d,e}

Traditional Chinese medicine (TCM), used in China and other Asian countries for thousands of years, is increasingly utilized in Western countries. However, due to inherent differences in how Western medicine and this ancient modality are practiced, employing the so-called Western medicine-based gold standard research methods to evaluate TCM is challenging. This paper is a discussion of the obstacles inherent in the design and statistical analysis of clinical trials of TCM. It is based on our experience in designing and conducting a randomized controlled clinical trial of acupuncture for post-operative dental pain control in which acupuncture was shown to be statistically and significantly better than placebo in lengthening the median survival time to rescue drug. We demonstrate here that PH assumptions in the common Cox model did not hold in that trial and that TCM trials warrant more thoughtful modeling and more sophisticated models of statistical analysis. TCM study design entails all the challenges encountered in trials of drugs, devices, and surgical procedures in the Western medicine. We present possible solutions to some but leave many issues unresolved. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: traditional Chinese medicine (TCM); acupuncture; postoperative dental pain; accelerated failure time model; blinding in randomized clinical trials

1. Introduction

Traditional Chinese medicine (TCM) is a coherent, well-developed system of medicine that has been practiced in China and neighboring Asian countries for thousands of years [1]. The system presents the human body as a whole and as a part of nature. Harmony must be maintained within body functions and between the body and nature in order to preserve health. Any disruption of this harmony leads to illness and disease. Over the millennia, ancient Chinese scholars and TCM practitioners elaborated and recorded unique diagnostic methods and treatment interventions, among them Chinese herbal medicine, acupuncture/moxibustion, *Tui Na* (Chinese massage and acupressure), mind/body exercise, and Chinese dietary therapy. Methods of disease prevention are an integral part of TCM and have become accepted by and widely popular among patients and medical communities in many Asian countries.

In recent years, complementary and alternative medicine (CAM), including TCM and acupuncture, has become increasingly popular in the U.S. According to a 2007 government survey, Americans spent \$14.8 billion out-of-pocket on non-vitamin, non-mineral natural products over the previous 12 months [2], a substantial increase from the \$5.1 billion spent in 1997. The same survey shows that patient visits to acupuncture clinics nearly tripled between 1997 (27.2/yr/1000) and 2007 (79.2/yr/1000). Regional and national surveys also show that the majority of physicians in the

^aCenter for Integrative Medicine, University of Maryland, School of Medicine, East Hall, 520 W. Lombard Street, Baltimore, MD 21201, U.S.A.

^bDepartment of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A.

^cCenter for Vaccine Development, University of Maryland, School of Medicine, 685 Baltimore St., Baltimore, MD 21201, U.S.A.

^dDivision of Biostatistics, Department of Epidemiology and Preventive Medicine, University of Maryland, School of Medicine, 10 S Pine St., Baltimore, MD 21201, U.S.A.

^eUniversity of Maryland Marlene and Stewart Greenebaum Cancer Center, 20 S Greene St., Baltimore, MD 21201, U.S.A.

*Correspondence to: Lixing Lao, Center for Integrative Medicine, University of Maryland, School of Medicine, East Hall, 520 W. Lombard Street, Baltimore, MD 21201, U.S.A.

†E-mail: llao@compmed.umm.edu

conventional medical practice consider acupuncture to be a legitimate medical modality [3–5], and in 1996, the U.S. Food and Drug Administration reclassified the acupuncture needle from a class III investigational device to a class II medical device (www.fda.gov). Research funding for CAM, including TCM and acupuncture, from the National Center for Complementary and Alternative Medicine (NCCAM), National Institute of Health (NIH) has significantly increased over the past 18 years, from a fiscal year budget of \$2 million in 1992, to \$50 million in 1999 when the NIH Office of Alternative Medicine became NCCAM, and to \$128.8 million in 2010 (<http://nccam.nih.gov/about/budget/appropriations.htm>, accessed 26 March 2010).

Despite these developments, the increased use of TCM in the West has raised efficacy and safety concerns among the medical community and the public [6–8]. Although numerous efforts have been made in the past decade to evaluate the effect of TCM, including acupuncture, true evaluation of this ancient treatment modality remains difficult and challenging, due to both the nature of its practice and its inherent differences from the Western medicine [9–11].

Scientific evaluation of this ancient healing art using such modern research methodology as the evidence-based medicine approach is warranted, and it has become urgent. A scientifically sound clinical trial based on the evidence-based approach to treatment evaluation entails two important aspects: (1) modern clinical trial design and (2) statistically sound methods of treatment effect assessment and inference. In Section 2 we use one of our randomized acupuncture clinical trials, a dental pain study (DPS), to illustrate how modern, evidence-based approaches can resolve some of the challenges of TCM research. Techniques to facilitate design of a valid randomized acupuncture clinical trial are highlighted in Section 3, and Section 4 presents the analytical aspects of treatment effect evaluation of acupuncture for pain control using statistical methods. Section 5 concludes this paper with a brief discussion.

2. Challenges in TCM research

An important and unique aspect of TCM practice is that it involves analysis of the status of the whole person using the so-called ‘holistic approach’ as opposed to a focus on a specific organ or disease. TCM diagnosis entails the ‘four diagnostic method,’ composed of four procedures: (1) inspection, (2) auscultation and olfaction, (3) inquiry, and (4) palpation [1], the aim of which are to collect and analyze information that reflects the condition of the body in order to determine its TCM syndrome(s). For example, according to the Western point of view, a disease such as osteoarthritis of the knee constitutes a single disease, determined by laboratory tests and standard criteria. From the perspective of TCM, this disorder may result from one or more TCM syndromes, for example ‘liver and kidney deficiency’, ‘cold- and dampness-induced disorder’, or ‘heat-induced disorder’, according to the patient’s condition. Syndromes are determined using the four diagnostic method and include data on the quality of sleep, appetite, bowel movements, and urination, among other information. TCM syndrome differentiation thus leads to individualized treatment plans that may require different treatment protocols [12] for a disorder that in Western medical practice warrants a single, standardized treatment.

2.1. Challenges in Chinese herbal research

Herbal medicine has a long history in both East and West, although traditionally Western practitioners tended to use ‘simples’ or single herbs, whereas Chinese practitioners have long used compounds of as many as 20 herbs. TCM methods require that Chinese herbal formulas be individualized according to the patient’s syndromes. But scientific research requires treatment standardization according to given Western medicine-defined diseases, which presents a problem for TCM herbal investigators. A possible solution is to use a formula that addresses a broad range of possible conditions that will benefit patients with a variety of TCM diagnoses, although this may compromise the effectiveness of the formula, as it may not be specific enough for the given TCM-diagnosed conditions.

An additional challenge in Chinese herbal research is the lack of standardization in herbal products, which is exacerbated by the fact that a single formula is composed of many herbs, each of which with numerous chemical compounds. These issues raise concerns regarding quality control and the consistency of herbal preparation, since quality and chemical constituents vary from field to field, season to season, and one extraction process to another. The solution to this problem is strict standards in sourcing, identification, compounding, biochemical testing, and quality control.

2.2. Challenges in acupuncture clinical research

Acupuncture has been used for millennia to treat a variety of diseases and symptoms in China. Recently, it has become increasingly popular in the U.S. as an adjunctive treatment to alleviate pain and reduce medication consumption, thus minimizing adverse drug effects [13, 14] in conditions such as back pain, headache, knee osteoarthritis, nausea, and vomiting due to chemotherapy, hypertension, and depression.

According to TCM theory, a vital energy known as *Qi* flows through a system of meridians and assists in homeostasis. Some 360 points distributed along the meridians serve both as diagnostic tools and as loci for acupuncture treatment [12]. It is presumed that when the normal flow of energy in the meridians is obstructed, for example as a result of tissue injury, pain or other symptoms result. Acupuncture treatment involves the insertion of small-gauge needles into specific acupuncture points for 15–30 min to restore the flow of vital energy through the affected meridians.

During the last decade, hundreds of clinical trials on acupuncture have been published. Despite these efforts to evaluate the effect of acupuncture, outcomes remain contradictory, largely due to problems in research design. The most challenging issue is blinding: it is not easy to keep patient and practitioner blinded. Patients are likely to feel the needle insertion, and practitioners are of course inserting the needles. In addition, unlike drug trials in which pills of identical size and color can be administered to patients in different treatment groups, acupuncture treatment is needle technique-dependent and can vary from practitioner to practitioner according to the length and quality of training and experience. In a single-blind trial, in which only the patients are blinded to group assignment, it is therefore crucial to select an adequate placebo control that can be used for blinding patients and is also inert, and, like a sugar pill, has no physiological effect. Furthermore, it is important to take TCM syndrome differentiation into consideration to yet meet the standard of research methodology by including a relatively homogeneous patient population. We addressed these key design issues in our study on acupuncture for dental pain control (DPS).

3. Design of the randomized clinical DPS trial

We conducted a randomized, placebo-controlled clinical trial using an innovative and scientifically sound research approach to evaluate the effect of acupuncture. The trial, entitled ‘Use of Acupuncture for Dental Pain: Testing a Model’, was conducted between 1998 and February 2002 at the Center for Integrative Medicine at the University of Maryland School of Medicine. It was funded by NCCAM, NIH (Grant #: 8 RO1 AT00010) and approved by the University of Maryland Institutional Review Board (IRB). The purpose of the trial was to evaluate the effect of acupuncture on the duration of post-operative surgical pain-free time in comparison to two sham acupuncture groups in a three-arm randomized controlled trial. Three-arm trials involving two control groups are quite unique and novel in acupuncture study.

The research methods have been previously reported [15]. In brief, 180 patients, 18–40 years old, of any race and all genders, who met the inclusive and exclusive criteria, which were that they be healthy and naïve to acupuncture, were recruited. After telephone screening by a study research coordinator, the patient was examined by a dental surgeon blinded to treatment assignment and, after completing the consent procedure and a pre-op questionnaire, was given an appointment for dental surgery. Each enrollee had one partially impacted bony lower third molar surgically removed by a board-certified oral and maxillofacial surgeon using 3 per cent carbocaine (range 54 to 150 mg) without vasoconstrictor. Immediately following surgery, the patient was randomly assigned to one of the three treatment groups ($n = 60$ per group): true acupuncture, sham insertion control (needles were inserted at ‘non-acupuncture points’), sham non-insertion placebo control (needles did not penetrate the skin). If patient reported moderate or severe pain, a second similar treatment was provided as part of the treatment protocol. Randomization was computer-generated in randomly permuted blocks of 3, 6, or 9 and was done via a secure call-in procedure to the coordinating center by the study acupuncturist immediately after surgery and just prior to the acupuncture treatment. For better design of control groups for treatment effect evaluation and to better address blinding in the randomized clinical trial—the biggest challenge in an acupuncture clinical trial, we carefully defined our three treatment groups and developed an intensive blinding mechanism to facilitate the trial procedure as follows:

Treatment Groups: The patient’s eyes were masked, and a licensed acupuncturist performed the treatment based on treatment assignment. For *true acupuncture*, the following acupuncture points were used: Hegu (LI 4), Jiache (St 6), Xiaguan (St 7), and Yifeng (SJ 17). One-in 32-gauge (0.25 mm) sterile, disposable needles were inserted 0.3–1.0 in. The needling sensation known as ‘De Qi’, soreness, heaviness, numbness, or distention, was obtained after insertion. The needles were retained for 20 min and then removed. For *sham acupuncture*, four non-acupuncture points (three local in the oral facial region and one in the hand) adjacent to the four points used in the real acupuncture group were used to constitute the sham study arm. Needles were inserted shallowly into these points and were not manipulated to avoid producing the ‘De Qi’ sensation. They were retained for 20 min, then removed. For *placebo acupuncture*, the points selected were identical to those of the true acupuncture group, but the needles were not inserted into the skin. To enhance blinding, a plastic needle tube was taped on the bony area next to each of the acupuncture points to produce a discernible sensation. A needle was taped to the dermal surface without skin penetration. Three manipulations (a dental instrument was used to probe the surface of the skin) were conducted every 10 min as in the treatment group. This placebo procedure for enhancing blinding was reported in our earlier studies [15].

Blinding and Masking: Patient blinding—Participants in the real acupuncture and control groups remained blind to assignment. A mask was used to occlude patients' eyes during the treatment sessions to facilitate blinding. The success of patient blinding was evaluated using a questionnaire after each acupuncture treatment and at the end of the on-site procedure. Patients were asked to indicate which kind of treatment they believed they had received (*acupuncture, sham/placebo, or don't know*) and the reason(s) that led to the belief. Successful patient blinding to treatment assignment was defined as a better-than-chance patient response of 'don't know'. **Research staff blinding**—Additionally, the dental surgeon, clinical coordinator, clinical assessor, and study investigators were blinded to patient assignment. The acupuncturists were the only study team members not blinded to patient assignment during the trial. However, the acupuncturists were blinded to specific patient information such as surgical evaluation or patient pain assessments.

Pain assessments were used as our primary endpoints. Immediately after the first treatment, the patient was asked to rate the pain level at 15-min intervals up to 6 h on site or until a rescue pain medication was given or requested. Pain assessments were reported using a questionnaire with two standardized pain scales: a four-point scale (0 = none, 1 = slight/mild, 2 = moderate, 3 = severe) and a 100 mm visual analog scale (VAS). The 'pain' event had to meet two criteria: both four-point scale ≥ 2 and VAS ≥ 30 mm. Although other pain assessments, including pain intensity difference (PID) and pain half gone (PHG), were collected, they were not used for primary outcomes. Our *primary outcomes* were: T_1 is the underlying true time to first pain (i.e. the true duration of first pain-free time) and T_2 is the underlying true time to rescue drug (i.e. the true duration of time before rescue drug). Both outcome measures were available from the questionnaires, since each patient had a timer in hand and recorded the minute when the first pain was felt and the pain level and the minute when a rescue drug was requested, although the data were presented in a longitudinal format. Even when patients did not feel pain, nurses reminded them to record their sensations for every 15 min. Of the 117 out of 180 patients who felt a first pain, 115 got a second treatment that was the same as the first; two patients refused the second treatment and requested rescue drug immediately. On-site pain assessment for patients who received a second treatment continued every 15 min for another 3 h following the second treatment or until a rescue pain medication was given or requested. By the end of the on-site study, a total of 102 patients out of 180 had requested the rescue drugs and 78 patients had not. Although patients continued to record their pain every hour until bedtime on the day of surgery after leaving the study site and then every day for 7 days following the surgery, the quality of self-reported data after leaving the study site is poor and the data were not well-collected. For this reason, we only focus on the on-site data in the current analysis.

The baseline covariates, which might be adjusted for this analysis, are age, gender, and race (five categories: African American, Asian, Caucasian, Hispanic, and others). Owing to the controlled randomized clinical trial design, it would be intuitive to use initial treatment group indicators in statistical analysis of our evaluation of the acupuncture effect. The challenge is that, according to the treatment protocol of the DPS (i.e. treatment could be repeated a second time if the patient felt pain after the first treatment and agreed a second treatment), only a selected subset (i.e. patients who felt pain) from each initial treatment group was chosen for the second treatment, which might threaten the validity of randomization for the treatment assignment mechanism. Although covariate adjustment is typically not needed for well-conducted and designed randomized clinical trials, this adjustment might help to control potential confounding effects from the residual imbalance of covariates across the three treatment groups after the second treatment. As a result, models with and without covariate adjustment are both shown in Section 4.

4. Statistical analysis of the DPS trial

Patients' baseline characteristics across the three treatment groups are summarized in Table I. Table I shows that there are no statistically significant differences in baseline characteristics among the randomized treatment groups, which confirms the validity of our initial randomization.

For our primary endpoints, time to first pain (T_1) and time to rescue drug (T_2), we only have right-censored data due to censoring mechanisms. Taking C_1 and C_2 as the censoring indicators for T_1 and T_2 , respectively, their relationship in the DPS is shown in Table II. The observed duration for first pain-free time is ' Y_1 ' = minimum (T_1 , censoring time for C_1) = minimum (T_1 , 360 min). The censoring mechanisms ($C_{1i} = 0$) could be that the patients remain 'pain free' (pain assessment did not meet both pain criteria) by the end of the study (the 6-h follow-up) or by the time of drop-out. Similarly, the observed duration of time before rescue drug ' Y_2 ' = minimum (T_2 , censoring time for C_2), where the censoring time is 360 min for patients without a second treatment and ($Y_1 + 180$) min for patients with a second treatment. The first two censoring mechanisms for C_2 are the same as the mechanisms for C_1 , i.e. if $C_{1i} = 0$ then $C_{2i} = 0$, which explains their dependent structure as shown in Table II. Additionally, for C_2 , a total of 15 patients were censored since they did not ask for rescue drug after the second treatment.

Based on those censoring mechanisms, a random censoring assumption is valid for the analysis of either (Y_1, C_1) or (Y_2, C_2) separately, but is not valid when we consider them jointly. Since this assumption is commonly used in standard

Table I. Summary of patients' baseline characteristics across three treatment groups.

Baseline characteristics		True acupuncture group	Sham acupuncture group	Placebo acupuncture group	P-value for global comparison
Age	Mean (SD)	25.50 (5.37)	24.58 (5.22)	26.32 (5.27)	0.1971*
Gender	Male	24	23	28	0.6188†
	Female	36	37	32	
Race	African American	18	14	15	0.8244†
	Asian	3	9	8	
	Caucasian	32	29	28	
	Hispanic	5	6	6	
	Others	2	2	3	

*F test based on ANOVA.

†Chi² test.

Table II. The relationship between two censoring indicators of the primary outcomes.

Table 1. Patient flow diagram													
Group	True acupuncture				Sham acupuncture			Placebo acupuncture			Total		
C2		0	1	Total	0	1	Total	0	1	Total	0	1	Total
C1	0	25	0	25	19	0	19	19	0	19	63	0	63
	1	6	29	35	8	33	41	1	40	41	15	102	117
Total		31	29	60	27	33	60	20	40	60	78	102	180

statistical software for survival analysis, both exploratory and model inferences for the separate analysis of (Y_1 , C_1) and (Y_2 , C_2) are highlighted in this paper. The joint analysis of multiple events allowing dependent censoring requires advanced statistical techniques and is not the focus of this paper.

For exploratory purposes, both Kaplan–Meier survival curves and log–log survival curves across acupuncture treatment groups are shown in Figure 1. And Table III compares the summary of non-parametric estimates of those survival curves for three treatment groups, together with the log-rank test results, for both Y_1 (the observed duration of first pain-free time) and Y_2 (the observed time to rescue drug).

For Y_1 , Table III shows that both the median and the mean durations of first pain-free time (in minutes) are the longest in the true acupuncture group, intermediate in the sham group, and the shortest in the placebo acupuncture group. Additionally, Table II shows that 25 patients remained pain free after the initial true acupuncture treatment until they were censored. In comparison with the 19 patients that remained pain free in both the sham and the placebo groups, the true acupuncture group seems to perform better in increasing the duration of post-operative surgical pain-free time. Figure 1 shows the comparisons of Kaplan–Meier curves between the true acupuncture group (group 'X') and the sham group (group 'Y') in (A1) and between the true group (group 'X') and the placebo group (group 'Z') in (A2). These further suggest the better performance of the acupuncture group, although this finding is not statistically significant based on the log-rank test, as shown in Table III.

For Y_2 , Table III shows that both the median and the mean duration of time before rescue drug are the longest in the true acupuncture group, intermediate in the sham group, and the shortest in the placebo acupuncture group. The log-rank test confirms that the true acupuncture group performed statistically significantly better in lengthening the duration of time to rescue drug for post-operative surgical pain control than did the placebo group. Additionally, Table II shows that 31 patients did not request rescue drug until they were censored in comparison to 27 patients in the sham and 20 in the placebo group. Meanwhile, a comparison of the well-separated Kaplan–Meier curves in (B1) of Figure 1 for the true acupuncture group (group 'X', top survival curve) and the placebo group (group 'Z', bottom survival curve) further confirms this finding.

For quantifying the acupuncture treatment effect on lengthening the duration of time to pain or rescue drug, the two most useful models are the Cox proportional hazard (Cox PH) model and the accelerated failure time (AFT) model [16]. The widely used Cox model is a semi-parametric model assuming constant treatment effect on the hazard (rate) ratio scale [17], whereas the less commonly used AFT model is a parametric model assuming the constant treatment effect on the survival time ratio scale [18, 19]. In comparison to the typical Cox PH model for our randomized DPS, i.e. $h(t) = \lambda_0(t) \exp(\beta_0 + \beta_1 \text{GroupX} + \beta_2 \text{GroupY})$, AFT models have the following form, without covariate adjustment:

$$\log(t) = \beta_0 + \beta_1 \text{GroupX} + \beta_2 \text{GroupY} + \sigma \varepsilon \quad (1)$$

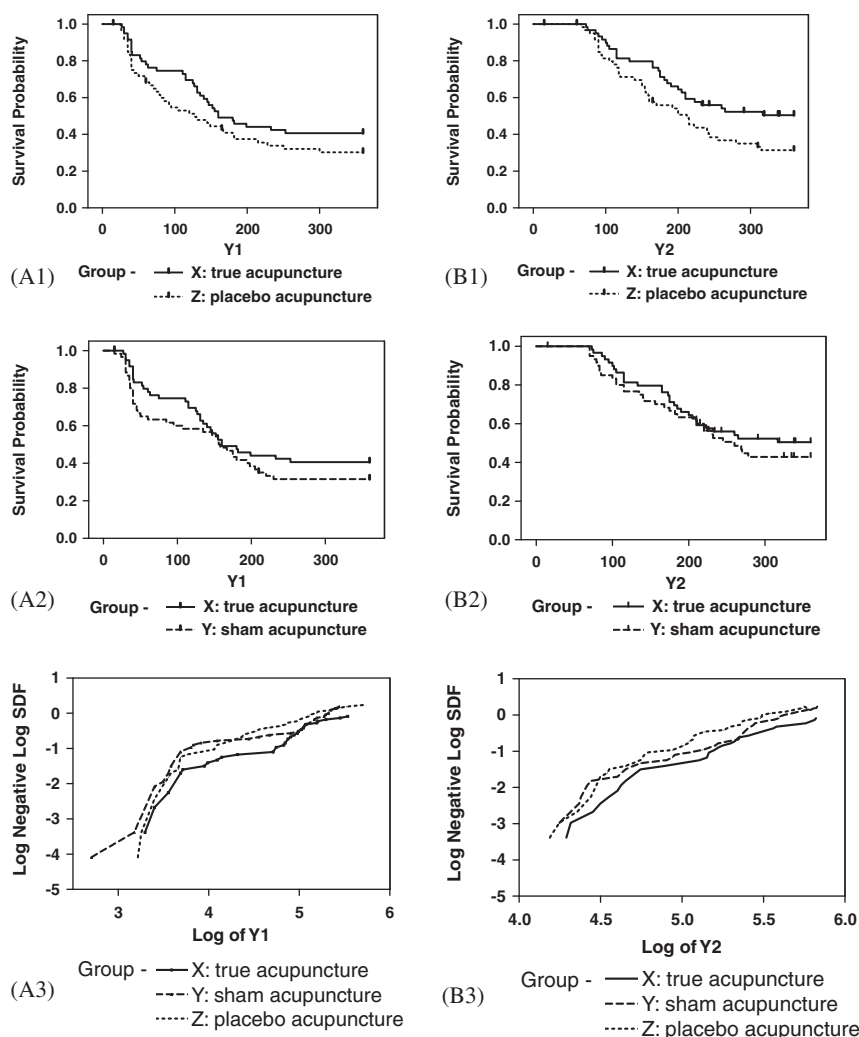


Figure 1. Kaplan–Meier curves and log–log curves across acupuncture treatment groups: (A1) K–M survival curves for the duration of first pain-free time (Y_1) across true acupuncture group (group ‘X’) vs placebo acupuncture group (group ‘Z’); (B1) K–M survival curves for the time to rescue drug (Y_2) across true acupuncture group (group ‘X’) vs placebo acupuncture group (group ‘Z’); (A2) K–M survival curves for the duration of first pain-free time (Y_1) across true acupuncture group (group ‘X’) vs sham acupuncture group (group ‘Y’); (B2) K–M survival curves for the time to rescue drug (Y_2) across true acupuncture group (group ‘X’) vs sham acupuncture group (group ‘Y’); (A3) Log[–log(estimated survival)] vs log(survival time) curves for the duration of first pain-free time (Y_1) across three groups (black = ‘X’, red = ‘Y’, and green = ‘Z’); and (B3) Log[–log(estimated survival)] vs log(survival time) curves for the duration time to rescue drug (Y_2) across three groups (black = ‘X’, red = ‘Y’, and green = ‘Z’).

where σ is the scale parameter and ε is the random error, GroupX and GroupY are the treatment indicators for the true and the sham acupuncture groups, respectively (with the placebo acupuncture group as the baseline), and β_1 and β_2 are the parameters of interest. The commonly used survival time distributions for AFT models are the Weibull, the exponential, the log-logistic, the log-normal, and the generalized gamma. To control the potential confounding effects of covariates, explained in Section 3, models both with and without covariate adjustment are fitted, and their results are summarized in Table IV.

The first two models for both outcomes in Table IV are the Cox PH models, due to the wide popularity of this model and the fact that it is semi-parametric. Table IV confirms our previous findings from Table III and Figure 1, and shows that: (1) the true acupuncture group performs statistically significantly better than does the placebo group in terms of lengthening the duration of time to rescue drug, and the hazard rate of getting rescue drug in the true acupuncture group is only 60 per cent (i.e. $e^{-0.504}$) that of the placebo group; (2) the true acupuncture group performs better than the placebo group in terms of reducing the hazard rate of getting first pain, but does not reach statistical significance. However, the cross-over phenomena in both log–log survival plots (A3, B3 in Figure 1) and the Kaplan–Meier curves (A2, B2 in Figure 1) indicate that the treatment effect (measured by hazard ratio) might not be constant but rather

Table III. Comparisons of estimated survival curves for both Y_1 (the observed duration of first pain-free time) and Y_2 (the observed time to rescue drug) across three acupuncture treatment groups.

	Treatment groups	N	Minimum	Maximum	Median	Mean	P value from Log Rank Test
Y_1	True	60	15	360	160	206.9	True vs Sham: 0.2507
	Sham	60	15	360	157	177.5	True vs Placebo: 0.1557
	Placebo	60	25	360	124	169.6	True vs (Sham+placebo): 0.1450
Y_2	True	60	15	360	338.5	259.2	True vs Sham: 0.4314
	Sham	60	70	360	252.5	245.6	True vs Placebo: 0.0338
	Placebo	60	60	360	200	217.7	True vs (Sham+placebo): 0.0997

Table IV. Evaluating acupuncture effects using the Cox Proportional Hazard (Cox PH) and accelerated failure time (AFT) models.

	Model	Treatment effects				Model selection criteria		
Y_1	<i>CoxPH models</i>	β_1	se	β_2	se	$-2\text{Log}(L)$	AIC	BIC
	M1	-0.314	0.230	-0.046	0.221	1110.2	1114.2	1119.8
	M2	-0.349	0.234	-0.062	0.224	1106.0	1122.0	1144.1
	<i>AFT models</i>							
	M1 (Log-logistic)	0.439	0.268	0.043	0.267	516.9	524.9	537.6
	M2 (Log-logistic)	0.442	0.271	0.068	0.268	513.3	533.3	565.3
	M1* (Log-normal)	0.402	0.257	0.014	0.253	508.4	516.4	529.2
	M2 (Log-normal)	0.401	0.257	0.029	0.254	505.2	525.2	557.1
Y_2	<i>CoxPH models</i>							
	M1	-0.504*	0.244	-0.297	0.235	975.4	979.4	984.6
	M2	-0.559*	0.247	-0.353	0.239	971.1	987.1	1008.0
	<i>AFT models</i>							
	M1 (Log-logistic)	0.362*	0.172	0.197	0.169	381.1	389.1	401.8
	M2 (Log-logistic)	0.392*	0.172	0.237	0.170	376.8	396.8	428.8
	M1* (Log-normal)	0.334*	0.166	0.168	0.164	373.2	381.2	393.9
	M2 (Log-normal)	0.355*	0.166	0.195	0.164	369.3	389.3	421.3

M1 are the corresponding CoxPH and AFT models without covariate adjustments, where M1* is the final model we choose for the DPS.

M2 are the corresponding CoxPH and AFT models adjusting for age, gender, and race.

AIC = Akaike's Information Criterion and BIC = Bayesian Information Criterion.

All models are fitted using PHREG and LIFEREG in SAS.

*Indicates that the corresponding P -value is less than 0.05.

become weakened over time. Consequently, the PH assumption might not hold for both survival outcomes, which casts doubt on the validity of inferences based on the Cox PH models, as shown in Table IV. This motivated us to pursue alternative AFT models.

We fitted AFT models with log-logistic, log-normal, and generalized gamma distributions, since these models do not have the proportional hazard property. AFT models under either Weibull or exponential distribution are not fitted, since they are equivalent to some parametric PH models under different parameterizations [19]. Furthermore, due to the questionable convergence criteria for the AFT model under generalized gamma distribution, only four models are reported in Table IV for each outcome based on log-logistic and log-normal distributions, including with (M2) and without (M1) covariate adjustment. For model selection among different distributions for each outcome, both Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) indicate that the log-normal AFT model without covariate adjustment, the M1* model in Table IV, fits our data the best. For both outcomes, the estimations of acupuncture effect on post-operative surgical pain control for the DPS based on their final M1* models are shown in Table V.

Table V confirms our previous findings from Table III and Figure 1, and shows that the true acupuncture group performed better than did the placebo group in terms of controlling post-operative surgical pain ($p < 0.05$). True acupuncture treatment lengthened the median survival time to rescue drug (Y_2) by an estimated factor of 1.397 times with 95 per cent CI of [1.008, 1.933] in comparison to placebo. Also, Table V shows that the true acupuncture group performed better than did the placebo group in terms of lengthening the median duration of first pain-free time (Y_1) by an estimated factor of 1.495 times, but without statistical significance. Also, Table V indicates that the variability in duration of first pain-free time (Y_1) is larger than survival time to rescue drug (Y_2) after accounting for treatment effects, as shown by the estimates of scale parameters.

Table V. Acupuncture effect estimations in the DPS using AFT model with log-normal distribution.

	Variable	β	se	e^{β}	95 per cent CI of e^{β}	<i>P</i> -value
Y_1	True acupuncture group (GroupX)	0.402	0.257	1.495	[0.904, 2.472]	0.117
	Sham acupuncture group (GroupY)	0.014	0.253	1.014	[0.618, 1.664]	0.956
	Scale	1.325	0.095		[1.152, 1.524]	
Y_2	True acupuncture group (GroupX)	0.334*	0.166	1.397*	[1.008, 1.933]	0.044
	Sham acupuncture group (GroupY)	0.168	0.164	1.183	[0.858, 1.629]	0.305
	Scale	0.837	0.065		[0.719, 0.975]	

*Indicates that the corresponding *P*-value is less than 0.05.

Importantly, the treatment effect estimates in Table V can be interpreted at the population level as well as at the individual level. Although the true reason why the PH assumption breaks down at the population level remains unknown, one potential reason may be unobserved heterogeneity in the study population. To account for this heterogeneity, a frailty model under the AFT model formulation is appropriate. And such models with Gamma (or inverse-Gaussian)-distributed frailty have a nice collapsibility property: if the AFT assumption holds at the individual level, then it must also hold at the population level using Gamma (or inverse-Gaussian)-distributed frailty [18, 19]. Consequently, the treatment effect estimates in Table V can be interpreted at the individual level given the validity of such frailty model in our DPS study. Hence, another interpretation for the estimated acceleration factor shown in Table V is that an individual receiving true acupuncture treatment instead of the placebo can statistically significantly lengthen his/her median survival time to rescue drug (Y_2) by an estimated factor of 1.397 with a 95 per cent CI of [1.008, 1.933]. On the individual level, interpretation of the acupuncture effect on lengthening median duration of pain-free time (Y_1) is similar, but without statistical significance.

5. Discussion

In the present study, we present a novel method of overcoming some methodological challenges in acupuncture clinical research. (1) We have successfully randomized the patients to achieve a balance among three groups in terms of gender, age, race, degree of surgical trauma, and local anesthetic usage. (2) We have successfully blinded the patients, who were not able to guess their treatment assignment correctly. (3) We were able to take into consideration various TCM diagnoses by using a homogenous patient population of healthy subjects without other medical conditions. As a result, the symptoms and TCM diagnosis were relatively the same, which minimized variation among the patients and meant that the same treatment protocol could be used for all.

Placebo-controlled RCTs are particularly difficult to conduct in China where many have grown up with acupuncture and are familiar with its sensations. Two factors may make such study possible there. (1) Although acupuncture has been used by many adults in China, the younger generation that will be enrolling in studies today is not a population accustomed to visiting acupuncture clinics. Thus, it may not be difficult to recruit acupuncture-naïve patients there. (2) With the recent developments in the acupuncture needle, needles are thinner than they used to be. With a guiding tube, they cause minimal sensation even in real acupuncture treatment. This may better blind patients in a real acupuncture group who expect a strong needle sensation as part of their treatment.

From the perspective of clinical trial design, TCM trials present the same challenges as those encountered in trials of drugs, devices, and surgical procedures. But as we demonstrate, data from TCM trials may have unique features that make them different from trials on other treatment modalities. Western medicine is based on disease specificity, laboratory indications, and standard criteria and treatment; TCM, on syndrome differentiation, the four diagnostic method, and individualized diagnosis and treatment. These differences pose difficulties for TCM research design and necessitate thoughtful consideration if challenges in statistical analysis are to be overcome and valid statistical inferences are to be derived. We have discussed various problems that might lead to the violation of the proportional hazard assumption. Such violation may explain why there are discrepancies between the analyses presented here and those presented in an earlier report [15].

If we understand the problems and account for them in the modeling, the statistical inference of treatment effect and estimation of the size of the treatment effect can be improved further. For example, the violation of the PH assumption for Y_2 might be caused by variations in the time of the second treatment in our DPS. In principle, a Cox PH model on (Y_2, C_2) with time-varying covariate can be fitted to account for this [19]. Unfortunately, due to the special censoring mechanisms, a zero event (no need for rescue drug) was observed for patients in the true acupuncture group who had first pain but refused the second treatment. In another words, the observed DPS data do not provide enough information to differentiate between the effects of getting true acupuncture treatment once versus twice, which causes numerical

problems for model fitting. In planning future trials, we may need to account for such special features in sample size calculation. A trial of a larger sample size may be able to model the recurrent event data, which may in turn prompt further development of the statistical methodology.

Acknowledgements

Drs. Lixing Lao and Brian Berman were supported by NIH Grant #: 8 RO1 AT00010 and Drs Lixing Lao, Brian M Berman, and Ming Tan were supported in part by the NIH Grant 1PO1 AT002605. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of NIH. We wish to thank Dr Lyn Lowry for editorial assistance and Ms Serena Lao for reference help.

References

1. Lao L. Traditional Chinese medicine, Part III, Chapter 12. In *Essentials of Complementary and Alternative Medicine*, Jonas WB, Levin JS (eds). Lippincott Williams and Wilkins: Philadelphia, 1999.
2. Nahin R, Barnes P, Stussman BJ, Bloom B. Costs of complementary and alternative medicine (CAM) and frequency of visits to CAM practitioners: United States, 2007. *National Health Statistics Reports* 2009; **18**:1–15. Accessed online 26 March 2010.
3. Berman B, Singh BK, Lao L, Singh B, Ferentz K, Hartnoll S. Primary care physicians' attitudes towards complementary medicine. *Journal of the American Board of Family Physicians* 1995; **8**(5):361–366.
4. Berman BM, Singh BB, Hartnoll SM, Singh BK, Reilly D. Primary care physicians and complementary–alternative medicine: training, attitudes, and practice patterns. *Journal of the American Board of Family Physicians* 1998; **11**(4):272–281.
5. Kurtz ME, Nolan RB, Rittinger WJ. Primary care physicians' attitudes and practices regarding complementary and alternative medicine. *Journal of the American Osteopathic Association* 2003; **103**(12):597–602.
6. Sea B (ed.). *Acute Pain Management: Operative or Medical Procedures and Trauma (95-0642)*. U.S. Department of Health and Human Services, Agency for Health Care Policy and Research, Public Health Service: Rockville, 1994; 49–50.
7. Cupp M. Herbal remedies: adverse effects and drug interactions. *American Family Physician* 1998; **58**(5):1133–1140.
8. Chan E, Tan M, Xin J, Sudarsanam S, Johnson DE. Interactions between traditional Chinese medicines and Western therapeutics. *Current Opinion in Drug Discovery and Development* 2010; **13**:50–65.
9. Liu T. Role of acupuncturists in acupuncture treatment. *Evidence-based Complementary and Alternative Medicine* 2007; **4**(1):3–6.
10. Tang JL, Liu BY, Ma KW. Traditional Chinese medicine. *Lancet* 2008; **6**,372(9654):1938–1940.
11. Tang JL. Research priorities in traditional Chinese medicine. *British Medical Journal* 2006; **333**(7564):391–394.
12. Cheng XN. *Chinese Acupuncture and Moxibustion*. Foreign Language Press: Beijing, 1987; 345.
13. Barnes P, Bloom B. Complementary and alternative medicine use among adults and children: United States, 2007. *National Health Statistics Reports* 2008; **12**:1–24. Accessed online 26 March 2010.
14. Sun Y, Gan TJ, Dubose JW, Habib AS. Acupuncture and related techniques for postoperative pain: a systematic review of randomized controlled trials. *British Journal of Anaesthesia* 2008; **101**(2):151–160.
15. Lao L, Bergman S, Hamilton G, Langenberg P, Berman B. Evaluation of acupuncture for pain control after oral surgery. *Archives of Otolaryngology: Head and Neck Surgery* 1999; **125**(5):567–572.
16. Hernan MA, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* 2005; **14**:477–491.
17. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman and Hall: London, 1984.
18. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
19. Kleinbaum DG, Klein M. *Survival Analysis: A Self-learning Text* (2nd edn). Springer: New York, 2005.