ORIGINAL ARTICLE

Translation in Different Diagnostic Procedures—Traditional Chinese Medicine and Western Medicine

Chin-Fu Hsiao,[1] Hsiao-Hui Tsou,[1] Yuh-Jenn Wu,[2] Chien-Hsiung Lin,[3] Yeu-Jhy Chang[4]*

Recently, the modernization of traditional Chinese medicines (TCM) for treatment of patients with critical and/or life-threatening diseases has attracted much attention in the pharmaceutical industry. However, there exist essential differences in the evaluation of the efficacy and safety of a TCM as compared with a typical Western medicine (WM), even though they are for the same indication. Therefore, the modernization of a TCM should be based on a scientific evaluation of the safety and effectiveness of the TCM in terms of well-established quantitative criteria. We propose a study design to study the calibration and validation of the Chinese diagnostic procedure for evaluation of a TCM, with respect to a well-established clinical endpoint for evaluation of a WM. Statistical validation of such an instrument is essential to have an accurate and reliable clinical assessment of the performance of the TCM. Similar to the validation of a typical quality of life instrument, some validation performance characteristics such as validity, reliability, and ruggedness are considered. In this article, a design for validation of a standard quantitative instrument to be commonly employed for diagnosis of patient function/activity, performance, disease signs and symptoms, and disease status and severity based on Chinese diagnostic practice is proposed. Methods for statistical validation of the standard instrument are derived. More specifically, for validation of the TCM diagnostic instrument, we consider the following validation performance characteristics (parameters): validity (or accuracy), reliability (or precision), and ruggedness (interrater variability). A numerical example is given to illustrate the proposed methods for validation of the Chinese diagnostic procedure. [*J Formos Med Assoc* 2008;107(12 Suppl):S74–S85]

Key Words: calibration, reliability, validation, validity

Recently, interest in alternative and complementary medicine has been growing in pharmaceutical research and development. In particular, many pharmaceutical companies have begun to focus on the modernization of traditional Chinese medicines (TCMs). With a history of over 3000 years, TCM is a natural and holistic medical system encircling the entire scope of human experience. It combines the use of Chinese herbal medicines, acupuncture, massage, and therapeutic exercise (e.g. Qigong, the practice of internal "air", and Taigie) for both treatment and prevention of disease. With its unique theories of etiology, diagnostic systems, and abundant historical literature, TCM itself consists of Chinese culture and philosophy, clinical practice experiences, and materials

[1]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, [2]Department of Applied Mathematics, Chung Yuan Christian University, Chung Li, [3]Department of Acupuncture and Moxibustion, Center for Traditional Chinese Medicine, and [4]Stroke Center and Department of Neurology, Chang Gung Memorial Hospital and College of Medicine, Chang Gung University, Taoyuan, Taiwan.

Received: July 25, 2008
Revised: September 24, 2008
Accepted: October 9, 2008

***Correspondence to:** Dr Chin-Fu Hsiao, Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan.
E-mail: chinfu@nhri.org.tw

including usage of many medical herbs. TCM drug treatment is typically comprised of complicated prescriptions of a combination of a few components. The combination is based on the Chinese diagnostic procedure (CDP).

It should be recognized that Western and Chinese medicine vary considerably even when they are used for the same indication.[1,2] Western doctors will first identify the cause and nature of disease, and treat patients accordingly, while Chinese doctors treat patients based on so called pattern discrimination. Experienced Chinese doctors believe that all of the organs within a healthy subject should reach the so-called global dynamic balance or harmony. Once the global balance is broken at certain sites such as heart, liver or kidney, some signs and symptoms then appear to reflect the imbalance at these sites. With respect to medical practice, we tend to see the therapeutic effect of Western medicines (WMs) faster than for TCMs. TCMs are often considered for patients who have chronic diseases or non-life-threatening diseases. For critical and/or life-threatening diseases, TCMs are gaining recognition as an alternative treatment.

Furthermore, the traditional CDP for a TCM is quite different from that of a WM. In general, the CDP consists of four major categories, namely, inspection, auscultation and olfaction, interrogation, and pulse taking and palpation. Chinese prescription of medicines then depends on a pattern that is derived from collecting symptoms and signs through these four diagnostic techniques. Inspection involves observing the patient's general appearance (strong or weak, fat or thin), mind, complexion (skin color), five sense organs (eye, ear, nose, lip, tongue), secretions, and excretions. Auscultation involves listening to the voice, expression, respiration, vomiting and coughing. Olfaction involves smelling the breath and body odor. Interrogation involves asking questions about specific symptoms and the general condition including history of the present disease, past history, personal life history, and family history. Pulse taking and palpation can help to judge the location and nature of a disease according to changes in the pulse. The smallest detail can have a strong impact on the treatment scheme as well as on the prognosis. Each category consists of a number of questions to collect different information regarding patient activity/function, disease status and/or disease severity. For example, the CDP for stroke consists of wind syndrome (six categories), fire–heat syndrome (nine categories), sputum syndrome (seven categories), stasis syndrome (five categories), deficiency syndrome (eight categories), and overabundant syndrome (nine categories), while WM uses the NIH Stroke Scale (NIHSS) developed by the US National Institute of Neurologic Disorder and Stroke (NINDS), from the original scale devised at the University of Cincinnati to measure the neurologic impact of stroke.[3]

The CDP is an instrument (or questionnaire) that consists of a number of questions designed to capture information that helps to determine the syndrome and/or condition to be treated. As a result, the CDP may be subjective. Consequently, the modernization of a TCM should be based on scientific evaluation of the efficacy and safety of the TCM in terms of well-established clinical endpoints for a Western indication through clinical trials on humans. When planning a clinical trial, it is suggested that the study objectives should be clearly stated in the study protocol. In practice, each clinical trial must have a primary question. At the design stage of a clinical trial, it is encouraged that the primary question should be carefully selected, clearly defined and stated in the study protocol. Once the primary question is identified, a valid study design can be chosen and the primary clinical endpoint can be determined accordingly. Based on the primary clinical endpoint, sample size required for achieving a desired power can then be calculated. For evaluating the therapeutic effect of a TCM, however, the commonly used clinical endpoint is usually not applicable, since the CDP may be subjective, as described in the previous section. As required by most regulatory agencies, such a subjective instrument must be validated before it can be used for assessment of treatment effect in clinical trials. As a result,

two questions may arise from the use of a CDP. First, it is of interest to determine the accuracy and reliability of this subjective diagnostic procedure for evaluation of patients with certain diseases. Second, it is also of interest to determine how a change of an observed unit in the CDP can be translated to a change in a well-established clinical endpoint for a Western indication. In this study, we addressed these two questions and proposed a study design to study the calibration and validation of the CDP for evaluation of a TCM with respect to a well-established clinical endpoint for evaluation of a WM. A numerical example is given to illustrate the proposed methods.

Materials and Methods

The CDP may be subjective. Therefore, it must be validated before it can be used for assessment of treatment effects in clinical trials, as required by most regulatory agencies. However, without a reference marker, not only can the CDP not be validated, but we do not know whether the TCM has achieved a clinically significant effect at the end of the clinical trial. Therefore, before the CDP for evaluation of a TCM can be validated with respect to a well-established clinical endpoint for evaluation of a WM, a calibration between the scale obtained from the CDP and the well-established clinical endpoint is necessary. Here, we propose a study design that allows calibration and validation of a CDP with respect to a well-established clinical endpoint for WM (as a reference marker). Subjects will be screened based on the criteria for Western indications. Qualifying subjects will be diagnosed by the CDP to establish a baseline. The subjects will then be randomized to receive either the test TCM or an active control (a well-established WM). Participating physicians including Chinese and Western clinicians will also be randomly assigned to either the TCM or WM arm. More specifically, this study design will result in three groups: Group 1—subjects who receive a WM, evaluated by a Chinese doctor and a Western clinician; Group 2—subjects who receive a TCM, evaluated by Chinese doctor A; Group 3—subjects who receive a TCM, evaluated by Chinese doctor B.

Group 1 can be used to calibrate the CDP against the well-established clinical endpoint, while Groups 2 and 3 can be used to validate the CDP based on the established standard curve for calibration. Based on the calibration model, a detected difference by the CDP can be translated to the well-established clinical endpoint. In addition, the CDP can also be validated against the well-established clinical endpoint. For validation of the TCM diagnostic instrument, we will consider the following validation performance characteristics (parameters): validity (or accuracy), reliability (or precision), and ruggedness (interrater variability).

Results

Calibration

Let N be the number of patients collected in Group 1. For the data in Group 1, let x_j be the measurement of the well-established clinical endpoint of the j^{th} patient for a WM. Suppose that the TCM diagnostic procedure consists of K items. Let z_{ij} denote the TCM diagnostic score of the j^{th} patient from the i^{th} item, $i = 1, \ldots, K$, $j = 1, \ldots, N$. Let y_j represent the score of the j^{th} patient summarized from the K TCM diagnostic items. For simplicity, we assume that

$$y_j = \sum_{i=1}^{K} z_{ij}.$$

For the data in Group 1, let x_j be the measurement of the well-established clinical endpoint of the j^{th} patient for a WM. For TCM instrument calibration, we consider two situations: the measurement of the well-established clinical endpoint is normally distributed; and the measurement of the well-established clinical endpoint is dichotomous. Here, the calibration should be performed for both baseline measurements and the measurements after treatment, since the relationship between y and x might be affected by the effect of medication.

The well-established clinical endpoint is normally distributed

Based on these measurements of WM clinical endpoints (standards) and their corresponding TCM scores, an estimated calibration curve can be obtained by fitting an appropriate statistical model between these standards and their corresponding TCM scores. The estimated calibration curve is also known as the standard curve. In a similar manner to that of calibration of an analytical method,[4,5] we will consider the following four candidate models:

- Model 1: $\gamma_j = \alpha + \beta x_j + e_j$,
- Model 2: $\gamma_j = \beta x_j + e_j$,
- Model 3: $\gamma_j = \alpha x_j^{\beta} e_j$,
- Model 4: $\gamma_j = \alpha e^{\beta x_j} e_j$,

where α and β are unknown parameters and the e values are independent random errors with $E(e_j) = 0$ and finite $Var(e_j)$ in Models 1 and 2, and $E(\log(e_j)) = 0$ and finite $Var(\log(e_j))$ in Models 3 and 4.

Model 1 represents a simple linear regression model which is the most commonly used statistical model for establishment of standard curves for calibration. Model 1 reduces to Model 2 when the standard curve passes through the origin. Models 3 and 4 are useful when there is a nonlinear relationship between γ and x. It should be noted that both Models 3 and 4 are in fact equivalent to a simple linear regression model after logarithmic transformation. For a given data set observed from Group 1, the standard curve under each model can be obtained by estimating the corresponding parameters through the least squares method. The standard curve is then used to evaluate the unknown WM clinical endpoint x_0 for a given TCM score γ_0. The unknown WM clinical endpoint is determined by solving x based on the standard curve, which assumes the parameter estimates are the true values of the parameters.

The well-established clinical endpoint is dichotomous

Suppose the measurement of the well-established clinical endpoint is either $x = 0$ or $x = 1$. In other words, x can be thought of as a classification variable defining groups of observations. For the calibration of the TCM diagnostic procedure, we can develop a discriminant criterion to classify each observation into one of the two groups ($x = 0$ or $x = 1$) based on the K TCM diagnostic items. Let $z_j = (z_{1j}, \ldots, z_{Kj})'$ be the K TCM diagnostic score of j^{th} patient. Assume that z_j for each group has a multivariate normal distribution. Based on the observed results from Group 1, we can derive the posterior probability of z_j belonging to group x, $p(x \mid z_j)$. The derivation of $p(x \mid z_j)$ is given in Appendix I. Consequently, an observation z_j is classified into group x if

$$p(x \mid z_j) = \max_{u=0,1} p(u \mid z_j).$$

When no assumptions can be made about the distribution within each group, or when the distribution is assumed not to be multivariate, nonparametric methods can be used to estimate the group-specific densities.[6,7]

Validity

For the sake of convenience, we assume that the well-established clinical endpoint is normally distributed. For the validity of a TCM instrument, we can evaluate the bias of the TCM instrument. That is, we are concerned about the accuracy of the TCM instrument, i.e. whether the questions in the TCM instrument are the right questions to capture the information regarding patient activity/function, disease status, and disease severity. We will use Group 2 to validate the CDP based on the previously established standard curve for calibration from Group 1. Let X be the unobservable measurement of the well-established clinical endpoint for WMs, which can be quantified by the TCM items, Z_i, $i = 1, \ldots, K$, based on the estimated standard curve in the previous section. Since both Models 3 and 4 can be transformed into a linear model using a log-transformation, for convention, we simply choose a linear model to illustrate the proposed methods for validation of the CDP. That is, we consider that

$$X = (Y - \alpha)/\beta,$$

where $Y=\sum_{i=1}^{K} Z_i$. That is, Model 1 was used for calibration. Suppose that X is distributed as a normal distribution with mean θ and variance τ^2. Let $Z=(Z_1, \ldots, Z_K)'$. Again, suppose Z follows a distribution with mean $\mu=(\mu_1, \ldots, \mu_K)'$ and variance Σ. To assess the validity, it is desired to see whether the mean of Z_i, $i=1, \ldots, K$ is close to $(\alpha+\beta\theta)/K$. Let $\bar{\mu}=\frac{1}{K}\sum_{i=1}^{K}\mu_i$. Then $\theta=(\bar{\mu}-\alpha)/\beta$. Consequently, we can claim that the instrument is validated in terms of its validity if

$$|\mu_i-\bar{\mu}|<\delta, \forall i=1,\ldots,K, \quad (1)$$

for some small prespecified δ. More specifically, to verify (1), it is desired to test the null hypothesis

$$H_0: |\mu_i-\bar{\mu}|\geq\delta \text{ for at least one } i. \quad (2)$$

To apply the approach of two one-sided tests, for each i, we will construct a $(1-\alpha)100\%$ confidence interval, (η_{i-}, η_{i+}), for $\mu_i-\bar{\mu}$. The construction for (η_{i-}, η_{i+}) is given in Appendix II. For each fixed i, a size α test based on the two one-sided tests approach rejects the hypothesis that $|\mu_i-\bar{\mu}|\geq\delta$ if and only if (η_{i-}, η_{i+}) is within $(-\delta, \delta)$. Then, using the approach of intersection–union, a size α test rejects the null hypothesis (2) and concludes that the TCM instruments are validated if and only if (η_{i-}, η_{i+}) is within $(-\delta, \delta)$ for all i.

Reliability

The calibrated well-established clinical endpoints derived from the estimated standard curve are considered reliable if the variance of X is small. We can now test the hypothesis

$$H_0: \tau^2\geq\Delta \text{ vs. } H_A: \tau^2<\Delta, \quad (3)$$

for some fixed Δ to verify the reliability of estimating θ by X. We will use Group 2 to verify the reliability based on the previously established standard curve for calibration. According to Lehmann,[8] we can construct a $(1-\alpha)100\%$ one-sided confidence interval for τ^2, say $(0, \xi)$. The calculation of ξ is given in Appendix III. Consequently, we can reject the null hypothesis (3) and conclude that the items are reliable in estimation of θ if $\xi<\Delta$.

Ruggedness

An experienced Chinese doctor usually prescribes a TCM based on the combined information obtained from the four major categories and his/her best judgment. In practice, the diagnostic procedure for a TCM can vary from one Chinese doctor to another. Although it may reduce within-patient variability, it can increase the between-rater variability, which can significantly bias the evaluation of the efficacy and safety of the TCM under study. Therefore, an acceptable TCM diagnostic instrument should produce similar results for different raters. In other words, it is desirable to quantify the variation caused by rater and the proportion of interrater variation to the total variation. We will use the one-way random model to evaluate instrument ruggedness.[4] A model describing a one-way random model is

$$x_{ij}=v+A_i+e_{ij},\ i=1 \text{ (Group 2), 2 (Group 3);}$$
$$j=1, \ldots, N,$$

where x_{ij} is the calibrated well-established clinical endpoint of the j^{th} patient obtained from the i^{th} rater derived from the estimated standard curve, v is the overall mean, A_i denotes the effect of the i^{th} rater and is assumed to be distributed i.i.d. $N(0, \sigma_A^2)$, and e_{ij} denotes the random error of the j^{th} patient's scale derived from the i^{th} rater, which is assumed to be distributed i.i.d. $N(0, \sigma_A^2)$. It is also assumed that A_i and e_{ij} are independent variables.[9]

To show that the interrater variability is within an acceptable limit ω, we can test the hypothesis

$$H_0: \sigma_A^2\geq\omega \text{ vs. } H_1: \sigma_A^2<\omega. \quad (4)$$

Since there exists no exact $(1-\alpha)100\%$ confidence interval for σ_A^2, we can then derive the Williams–Tukey interval,[10] (L_A, U_A), with a confidence level between $(1-2\alpha)100\%$ and $(1-\alpha)100\%$ for σ_A^2. The derivation of (L_A, U_A) is shown in Appendix IV. Accordingly, the null hypothesis (4) is rejected at the α level of significance if $U_A<\omega$.

Numerical example

To illustrate the methods proposed, a randomized trial was conducted to study the effect of acupuncture on stroke patients. Patients with an acute ischemic stroke between 4 and 10 days were

allocated into three groups. The diagnostic criteria of acute ischemic stroke consisted of the typical presentations of acute onset of focal neurologic deficits, and excluded other possible organic brain lesions by brain computed tomography and/or magnetic resonance imaging. Thirty stroke patients received aspirin 100 mg/day and were evaluated by a Chinese doctor and a Western clinician (Group 1), 30 stroke patients received acupuncture and were evaluated by Chinese doctor A (Group 2), and 30 stroke patients received acupuncture and were evaluated by Chinese doctor B (Group 3). The combination of scalp and body acupoints that fit the Chinese traditional theory was applied in patients from Groups 2 and 3. The measurement that the Western clinician used was the NIHSS, whereas the TCM diagnostic instruments considered in this study were wind and fire–heat syndromes. More specifically, patients in Group 2 had both NIHSS and TCM scores, while patients in Groups 2 and 3 had only TCM scores. Outcome assessments were recorded at randomization, 14 days, 1 month, 3 months, and 6 months after treatment.

The TCM instruments are summarized based on the rating scales of the wind and fire–heat syndromes shown in Table 1, that is, $K = 2$. More specifically, the wind syndrome is a rating scale with six categories, including onset conditions (0–8), limb condition (0–7), tongue body (0–7), eyeball condition (0–3), string-like pulse (0–3), and head condition (0–2). Patients with a total score > 7 were considered to have wind syndrome. On the other hand, the fire–heat syndrome consists of nine categories, including tongue condition (0–6), tongue fur (0–5), stools (0–4), spirit (0–4), facial and breath conditions (0–3), fever (0–3), pulse (0–2), mouth (0–2), and urine (0–1). Again, patients with a total score > 7 were considered to have fire–heat syndrome. In both syndromes, the larger the scale, the more severe the syndrome. Data are shown in Tables 2 and 3.

Let γ represent the sum of the scores of wind and fire–heat syndromes and x represent the NIH stroke score. Here, we used the baseline measurements for calibration. From Group 1, the estimated standard curve based on Model 1 was given as $\gamma = 7.092 + 1.820x$. The estimated regression line and the original data are presented in the Figure.

Group 2 was used to validate the CDP based on the previously established standard curve. We claimed that the instruments of wind and fire–heat syndromes were validated if

$$|\mu_i - \bar{\mu}| < \delta, \forall i = 1,\ 2,$$

for some small prespecified δ. It can be seen from Group 2 that $\hat{\mu}_1 = 9.733$ and $\hat{\mu}_2 = 7.067$.

Accordingly, (η_{1-}, η_{1+}) and (η_{2-}, η_{2+}) were respectively given by (0.328, 2.338) and (−2.338, −0.328). In this case, we could reject the null hypothesis (2) if $\delta = 3$.

We also used Group 2 to evaluate the reliability of the items for the TCM instrument. That is, the wind and fire–heat syndromes for the TCM instrument were considered reliable if the variance of X derived from the previously established

Table 1. Wind and fire–heat syndromes

Wind syndrome		Fire–heat syndrome	
Category	Score	Category	Score
Onset conditions	0–8	Tongue conditions	0–6
Limb conditions	0–7	Tongue fur	0–5
Tongue body	0–7	Stool	0–4
Eyeball conditions	0–3	Spirit	0–4
String-like pulse	0–3	Facial and breath conditions	0–3
Head conditions	0–2	Fever	0–3
		Pulse	0–2
		Mouth	0–2
		Urine	0–2

Table 2. Data for Group 1

Subject ID	TCM score (Wind + Fire–heat)	NIH stroke score
1	19	6
2	11	2
3	8	2
4	10	2
5	16	4
6	19	8
7	22	9
8	10	2
9	18	4
10	15	6
11	21	8
12	13	5
13	23	8
14	26	10
15	13	5
16	32	13
17	17	5
18	18	6
19	11	3
20	23	7
21	12	2
22	27	11
23	12	2
24	22	8
25	17	5
26	13	3
27	13	5
28	31	13
29	15	4
30	17	3

standard curve was small. Assume that $\Delta = 15$. From Group 2, a 95% one-sided confidence interval for τ^2 was (0, 13.48). Since 13.48 is < 15, we could reject the null hypothesis (3) at the 5% level of significance, and conclude that the TCM instrument was validated in terms of its precision. Selection of Δ should reflect the considerable information that existed in previous studies. It may also vary from disease to disease.

Groups 2 and 3 were used to quantify the variation caused by raters. The response variable was logarithmically transformed to normalize their distributions. The ANOVA is given in Table 4, which shows that SSA = 0.012 and SSE = 13.813. Hence, estimates for σ_A^2 and σ^2 were given by $\hat{\sigma}^2 = 0.012$ and $\hat{\sigma}_A^2 = 0$. Since $F = 0.05$ with a p value of 0.8262, we could not reject the null hypothesis $H_0: \sigma_A^2 = 0$ at the 5% level of significance. The Williams–Tukey interval with a confidence level between 90% and 95% for σ_A^2 was given by (0, 0.399). This suggests that the interrater variation was not significant.

Discussion

Although the modernization of TCM for treatment of patients with critical and/or life-threatening diseases has attracted much attention in the pharmaceutical industry, it should be recognized that there are fundamental differences in the scientific evaluation of the efficacy and safety of a TCM as compared with a typical WM. The validation of a standard quantitative instrument in a TCM clinical trial plays an important role in providing an accurate and reliable assessment of the safety and effectiveness of the TCM under investigation. Most importantly, the calibration of the quantitative instrument with respect to a well-established clinical endpoint provides clinicians with a better understanding of whether the observed significant difference from the quantitative instrument is clinically meaningful. It should be noted that only a well-calibrated and validated quantitative instrument is able to lead to accurate estimation of the sample size required for achieving a desired power for detecting a clinically meaningful difference.

In this study, four common statistical models were used for the calibration of the CDP with respect to a well-established clinical endpoint. However, the relationship between the CDP and the well-established WM clinical endpoint may vary considerably from disease to disease. For some diseases, the relationship might be linear. In some cases, a generalized linear model may be more suitable for the relationship between CDP and WM. Therefore, intensive research in the design and analysis method might be needed to correctly interpret the relationship between the CDP and WM. If the relationship between the TCM score

Table 3. Data for Groups 2 and 3

Group 2 subject ID	Wind	Fire–heat	Group 3 subject ID	Wind	Fire–heat
1	8	3	1	11	3
2	17	6	2	11	8
3	7	10	3	11	8
4	11	0	4	11	4
5	7	3	5	9	5
6	7	13	6	5	11
7	9	3	7	12	3
8	18	4	8	13	3
9	12	10	9	13	6
10	13	9	10	10	5
11	7	3	11	11	7
12	13	10	12	8	7
13	15	6	13	7	6
14	7	16	14	8	5
15	9	11	15	9	4
16	11	12	16	8	7
17	5	7	17	11	6
18	5	16	18	17	1
19	5	6	19	13	4
20	11	0	20	8	3
21	12	0	21	13	3
22	5	8	22	7	8
23	12	0	23	11	6
24	12	10	24	7	6
25	9	4	25	12	3
26	7	6	26	8	9
27	7	3	27	7	9
28	14	8	28	11	7
29	8	12	29	10	5
30	9	13	30	13	4

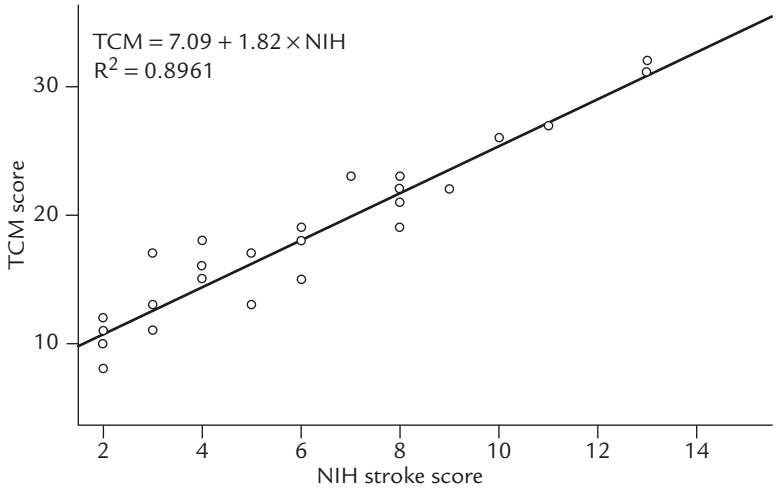

Figure. Scatter plot of sum of wind syndrome score and fire–heat syndrome score versus NIH stroke score for the data in Group 1, and the estimated standard curve.

Table 4. Analysis of variance table for data in Groups 2 and 3

Source of variation	Degrees of freedom	Sum of squares	Mean square	*F*	*p*
Rater	1	0.012	0.012	0.05	0.8262
Error	58	13.813	0.238		
Total	59	13.825			

and the WM endpoint is not one of the four candidate models, more complicated calibration functions or transformations may be required.

Note that in the example, we used the baseline measurements for calibration to illustrate our approach. However, it is strongly suggested that the calibration should be performed for baseline measurements and those after treatment, since the relationship between y and x might be affected by the effects of a medication. Also note that when larger variation caused by raters occurs, two questions may arise. First, the CDP instrument may be defective. Second, TCM doctors might have different TCM practices or experiences. For the former case, the CDP instrument needs to be refined. For the latter case, the rater should revisit the established Chinese diagnostic criteria in order to ensure that consistency is maintained.

We tend to believe that TCMs are mostly made of natural herbs, and thus are nearly free from side effects and much less toxic than Western drugs. However, scientific documentation regarding clinical evidence of safety and efficacy of these TCMs remain limited. Although the use of TCM in humans has a history of more than 3000 years, there have been no regulatory requirements with regard to the assessment of safety and effectiveness of TCMs until recently. However, the regulatory authorities of both China and Taiwan have now published guidelines for clinical development of TCMs.[11–13] In addition, the United States Food and Drug Administration has also published guidance for botanical drug products.[14] These regulatory requirements for TCM research and development, especially for clinical development, are very similar to the well-established guidelines for pharmaceutical research and development for WMs. It is unclear whether these regulatory requirements are feasible for the research and development of TCM given that there are so many fundamental differences in medical practice, drug administration and diagnostic procedures. Consequently, it is strongly suggested that current regulatory requirements should be modified to reflect these fundamental differences.

References

1. Chow SC, Pong A, Chang YW. On traditional Chinese medicine clinical trials. *Drug Inf J* 2006;40:395–406.
2. Tse SK, Chang JY, Su WL, et al. Statistical quality control process for traditional Chinese medicine. *J Biopharm Stat* 2006;16:861–74.
3. Lyden P, Lu M, Jackson C, et al. NINDS tPA Stroke Trial Investigators. Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. *Stroke* 1999;30:2347–54.
4. Chow SC, Liu JP. *Statistical Design and Analysis in Pharmaceutical Science*. New York: Marcel Dekker, 1995.
5. Tse SK, Chow SC. On model selection for standard curve in assay development. *J Biopharm Stat* 1995;5:285–96.
6. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann Math Stat* 1956;27:832–7.
7. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat* 1962;33:1065–76.
8. Lehmann EL. *Testing Statistical Hypotheses*, 2nd edition. New York: Wiley, 1986.
9. Searle SR, Casella G, McCulloch CE. *Variance Components*. New York: Wiley, 1992.
10. Williams JS. A confidence interval for variance components. *Biometrika* 1962;49:278–81.
11. Ministry of Public Health. *Guidance for Drug Registration*. Beijing, China: Ministry of Public Health, 2002.
12. Department of Health. *Draft Guidance for IND of Traditional Chinese Medicine*. Taipei: Department of Health, Taiwan, 2004.
13. Department of Health. *Draft Guidance for NDA of Traditional Chinese Medicine*. Taipei: Department of Health, Taiwan, 2004.
14. United States Food and Drug Administration. *Guidance for Industry—Botanical Drug Products*. Rockville, MD: US FDA, 2004.

Appendix I

Let m_x be the K-dimensional vector containing score means in group x. Let S_x and S_p denote the covariance matrix within group x and the pooled covariance matrix respectively. The squared Mahalanobis distance from z_j to group x can be expressed as

$$d_x^2(z_j)=(z_j-m_x)'V_x^{-1}(z_j-m_x),$$

where V_x can be chosen as S_x or S_p. Accordingly, the group-specific density estimate at z_j from group x is given by

$$f_x(z_j)=(2\pi)^{-K/2}|V_x|^{-1/2}\exp(-0.5d_x^2(z_j)).$$

Let q_x be the prior probability of membership in group x. By applying Bayes' theorem, the posterior probability of z_j to group x is given by

$$p(x\mid z_j)=\frac{q_x f_x(z_j)}{\sum_u q_u f_u(z_j)},$$

where the summation is over both groups. The generalized square distance from z to group x can be defined as

$$D_x^2(z_j)=d_x^2(z_j)+g_1(z_j)+g_2(z_j),$$

where

$$g_1(z_j)=\begin{cases}\ln|S_x| & \text{if } V_x=S_x,\\ 0 & \text{if } V_x=S_p,\end{cases}$$

and

$$g_2(z_j)=\begin{cases}-2\ln|q_x| & \text{if } q_0\neq q_1,\\ 0 & \text{if } q_0=q_1,\end{cases}$$

where $|S_x|$ is defined as the determinant of S_x. As a result, the posterior probability of z_j belonging to group x is equal to

$$p(x\mid z_j)=\frac{\exp(-0.5D_x^2(z_j))}{\sum_u \exp(-0.5D_u^2(z_j))}.$$

Appendix II

We can write

$$\mu_i-\bar{\mu}=a_i'\mu,\ i=1,\ldots,K,$$

where $a_i=\begin{pmatrix}-\frac{1}{K}1_{i-1}\\ 1-\frac{1}{K}\\ -\frac{1}{K}1_{K-i}\end{pmatrix}$, $1_i=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{(i-1)\times 1}$, and $1_{K-i}=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{(K-i)\times 1}$.

Assume that the TCM instrument is administered to N patients from Group 2. Let $\hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} Z_j = \bar{Z}$.

Consequently, we can derive that

$$\eta_{i\pm} = a_i'\hat{\mu} \pm t_{1-\alpha;\, N-1}\sqrt{\frac{1}{N}a_i' S a_i},$$

where $t_{1-\alpha;\, N-1}$ is the $(1-\alpha)^{\text{th}}$ quantile of the t-distribution with $N-1$ degrees of freedom.

Appendix III

Based on the estimated standard curve, we can derive that

$$\begin{aligned}\tau^2 &= \frac{1}{\beta^2}\mathrm{Var}\left(\sum_{i=1}^{K} Z_i\right)\\ &= \frac{1}{\beta^2}1'\sum 1.\end{aligned}$$

Note that the sample distribution of

$$\sum_{j=1}^{N}(X_j - \bar{X})^2/\tau^2$$

has a χ^2 distribution with $N-1$ degrees of freedom. According to Lehmann,[8] we can construct a $(1-\alpha)100\%$ one-sided confidence interval for τ^2 as follows

$$\begin{aligned}\tau^2 &\geq \frac{\sum_{j=1}^{N}(X_i - \bar{X})^2}{\chi^2(\alpha, N-1)}\\ &= \xi.\end{aligned}$$

Appendix IV

Two sums of squares are the sum of squares within, SSE, and the sum of squares between, SSA. That is,

$$\mathrm{SSE} = \sum_{i=1}^{2}\sum_{j=1}^{N}(x_{ij} - \bar{x}_{i\bullet})^2,$$

and

$$\mathrm{SSA} = N\sum_{i=1}^{2}(\bar{x}_{i\bullet} - \bar{x}_{\cdot\cdot})^2,$$

where $\bar{x}_{i\bullet} = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$ and $\bar{x}_{\cdot\cdot} = \frac{1}{2N}\sum_{i=1}^{2}\sum_{j=1}^{N} x_{ij} = \frac{1}{2}\sum_{i=1}^{2}\bar{x}_{i\bullet}$. Let MSA and MSE denote mean squares for factor A and mean square error. Then MSA = SSA and MSE = SSE/[2($N-1$)]. As a result, the analysis of variance estimators of σ^2 and σ_A^2 can be obtained as follows:

$$\hat{\sigma}^2 = \mathrm{MSE}$$

and

$$\hat{\sigma}_A^2 = \frac{\text{MSA} - \text{MSE}}{N}.$$

Consequently, the Williams–Tukey interval,[10] (L_A, U_A), with a confidence level between $(1-2\alpha)100\%$ and $(1-\alpha)100\%$ for σ_A^2 can be expressed as

$$L_A = \frac{\text{SSA}(1 - F_U / F_A)}{N\chi^2_{U_A}},$$

and

$$U_A = \frac{\text{SSA}(1 - F_L / F_A)}{N\chi^2_{L_A}},$$

where $F_L = F(1-0.5\alpha, 1, 2(N-1))$ and $F_U = F(0.5\alpha, 1, 2(N-1))$ represent the $(1-0.5\alpha)^{\text{th}}$ and $(0.5\alpha)^{\text{th}}$ upper quantiles of a central F distribution with 1 and $2(N-1)$ degrees of freedom, $\chi^2_{L_A} = \chi^2(1-0.5\alpha, 1)$ and $\chi^2_{U_A} = \chi^2(0.5\alpha, 1)$ are the $(1-0.5\alpha)^{\text{th}}$ and $(0.5\alpha)^{\text{th}}$ upper quantiles of a central χ^2 distribution with 1 degree of freedom, and $F_A = \text{MSA}/\text{MSE}$.